Coupled dynamics of sequence selection and compactification in mean-field hetero-polymers

H Chakravorty†**, A C C Coolen**† **and D Sherrington**‡

†Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK

‡Department of Physics – Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK

Abstract. We study a simple solvable model describing the genesis of monomer sequences for hetero-polymers (such as proteins), as the result of the equilibration of a slow stochastic genetic selection process which is assumed to be driven by the competing demands of functionality and reproducibility of the polymer's folded structure. Since reproducibility is defined in terms of properties of the folding process, one is led to the analysis of the coupled dynamics of (fast) polymer folding and (slow) genetic sequence selection. For the present mean-field model this analysis can be carried out using the finite-dimensional replica method, leading to exact results for (first- and second-order) transitions and to rich phase diagrams.

PACS numbers: 61.41.+e, 75.10.Nr

1. Introduction

The functionality of a protein depends largely on the shape of its 3D native state. Determining this shape from a protein's sequence of amino-acids is called the 'protein folding problem'. From a medical point of view it is desirable to be able to solve the related 'inverse folding problem': given a 3D structure, find a sequence of amino-acids which would have this as its native state. Unfortunately, X-ray crystallography (the standard tool for determining native states) is very time-consuming, and will not deliver the native states of the many amino-acid sequences collected through the Human Genome Project for years to come [1, 2]. Furthermore, it does not reveal the mechanics of the folding process. Molecular dynamics simulations are also extremely slow; only the first 10-50 ns of protein folding processes could so far be simulated [3, 4, 5, 6]. In spite of ongoing hardware improvements, it will take long before real progress is made. This points to the need for solvable mathematical models, designed to capture the essentials of proteins, to shed more light on the nature of the folding process and the role of its parameters.

The main complications in solving protein models are the chain constraint and the presence of disorder, embodied in the amino-acid sequences (see e.g. [7, 8]). Natural amino-acid sequences have evolved genetically, driven mainly by the two (competing) demands of structure reproducibility and functionality. They are not random (random hetero-polymers usually do not fold into unique shapes), whereas most disordered systems techniques are based on exploiting self-averaging properties of random disorder. It is not yet understood what distinguishes a random amino-acid sequence from a natural one. Our objective is to define and study a solvable model describing the interplay between the genetic forces and folding processes, as a first step towards understanding the genesis and statistics of natural amino-acid sequences.

Our model is a simple mean-field hetero-polymer whose microscopic state is described by two degrees of freedom per monomer: $\phi_i \in [0, 2\pi]$, giving the orientation of monomer i relative to the backbone, and $\eta_i \in \{-1, 1\}$, giving its polarity (i.e. hydrophobic vs. hydrophilic)‡. The evolution of the orientation variables represents the folding; that of the polarity variables (which involves changing monomer species) represents genetic evolution. The latter clearly takes place over much larger time-scales than the former, and is the result of an interplay between minimizing an energetic cost determined by the average folding quality, constraints on the monomer composition, and noise. Models with adiabatically separated time-scales have been studied and solved using finite-dimensional replica theories, but so far mainly in the context of neural networks and

‡ Thus, in addition to the mean-field nature of the forces and the absence of a chain constraint, our model simplifies biological reality further by reducing the orientation degrees of freedom of individual amino-acids to one, and the characterization of the physical properties of individual amino-acids to a binary number.

spin-glasses [9, 10, 11, 12, 13, 14, 15, 16]. We show how these techniques can also be used to study analytically the coupled dynamics of (fast) folding and (slow) genetic sequence selection in our present model.

We first define our model and solve it in the stationary state, describing equilibrium at the largest (genetic) time-scale, via the finite-dimensional replica method. Analysis of the resulting macroscopic equations leads to general results regarding the ground state and (first- and second-order) phase transitions. We inspect our equations more closely for specific (small) values of the number q of allowed local monomer orientations, and generate phase diagrams via a combination of analytical and numerical techniques. These are found to be very rich, and involve various types of single-state and multiple-state swollen and compact phases.

2. The model and its solution

2.1. Model definitions

We study hetero-polymer models as in [17], consisting of N monomers labelled $i = 1 \ldots N$. The spatial degrees of freedom of the monomers are angles $\phi_i \in C \subseteq [0, 2\pi]$, describing their orientations around a polymer backbone; each monomer i can take only a finite number of positions on a circle. The physical properties of a monomer i are defined by its species (i.e. amino-acid type). Here we restrict ourselves by taking into consideration only a monomer's polarity $\eta_i = \mathbb{R}$, where $\eta_i > 0$ for hydrophylic monomers and $\eta_i < 0$ for hydrophobic ones. A monomer sequence is thus fully described by the vector $\boldsymbol{\eta} = (\eta_1, ..., \eta_N) \in \mathbb{R}^N$, whereas the spatial configuration (or 'conformation') of the system is described by the vector $\boldsymbol{\phi} = (\phi_1, ..., \phi_N) \in C^N$.

On the time-scales of a single evolutionary generation the sequence $\boldsymbol{\eta}$ is fixed, and the only allowed process is folding, i.e. evolution of the $\boldsymbol{\phi}$. We consider here only the dominant folding force: compactification of the polymer via 'shielding' from the solvent of its hydrophobic monomers. A simple well-known type of phenomenological Hamiltonian to describe this effect is (see e.g. [18, 19, 20])

$$H(\boldsymbol{\phi}, \boldsymbol{\eta}) = -\frac{J}{N} \sum_{ij} \eta_i \eta_j \ \delta[\phi_i - \phi_j] \tag{1}$$

with $J > 0$. The rationale is that efficient shielding of hydrophobic monomers and exposure of hydrophilic ones requires *separation* of the species; ideally with the two polarity types oriented at opposite sides of the chain (here a single bend of the chain would ensure shielding). The Hamiltonian (1) punishes monomer pairs with opposite polarity but identical orientation relative to the backbone (which make polarity-based compactification more difficult), and favors pairs with identical polarity and orientation. The mathematical simplifications induced by the separable character and the infinite range of the monomer interactions in (1) will allow us to solve the super-imposed slow genetic process to be introduced below.

For a fixed realization of the sequence $\boldsymbol{\eta}$, the equilibrium statistics of the orientation variables, at temperature $T = \beta^{-1}$, are characterized by the partition function

$$Z[\boldsymbol{\eta}] = \mathrm{Tr}_{\boldsymbol{\phi}} \ e^{-\beta H(\boldsymbol{\phi}, \boldsymbol{\eta})} \tag{2}$$

In addition to the folding process we now introduce a stochastic dynamics for the sequences $\boldsymbol{\eta}$, which reflects the demands of structure reproducibility and structure functionality. The former is measured by the achieved degree of species separation, i.e. by (1). Upon describing the degree of functionality of a sequence $\boldsymbol{\eta}$ by a potential $V(\boldsymbol{\eta})$, we arrive at the Langevin equation

$$\frac{d}{dt}\eta_i = -\frac{\partial}{\partial \eta_i} \left\{ H(\boldsymbol{\phi}, \boldsymbol{\eta}) + V(\boldsymbol{\eta}) \right\} + \xi_i \tag{3}$$

with zero-average Gaussian random forces $\xi_i(t)$, obeying $\langle \xi_i(t) \xi_j(t') \rangle = 2\tilde{T} \delta_{ij} \delta(t - t')$ (which introduces a second 'temperature' $\tilde{T} = \tilde{\beta}^{-1}$). Our further discussion will be restricted to the stationary state of the genetic process (3). Since the genetic process is adiabatically slow compared to the folding process, we may replace (3) by its average over the equilibrium folding statistics $p(\boldsymbol{\phi}|\boldsymbol{\eta}) \sim e^{-\beta H(\boldsymbol{\phi}, \boldsymbol{\eta})}$, i.e.

$$\frac{\partial H(\boldsymbol{\phi}, \boldsymbol{\eta})}{\partial \eta_i} \quad \to \quad \langle \frac{\partial H(\boldsymbol{\phi}, \boldsymbol{\eta})}{\partial \eta_i} \rangle_{\boldsymbol{\phi}} = \frac{\partial}{\partial \eta_i} \left\{ -\frac{1}{\beta} \log Z[\boldsymbol{\eta}] \right\}$$

The equilibrium measure of the genetic process is now also of the Boltzmann form, with a 'genetic' Hamiltonian $H(\boldsymbol{\eta})$ which is the sum of the free energy of the folding process, given the sequence $\boldsymbol{\eta}$, and the functionality potential $V(\boldsymbol{\eta})$), and characterized by an associated genetic free energy per monomer f_N (with the 'folding partition function' defined in (2)):

$$H(\boldsymbol{\eta}) = V(\boldsymbol{\eta}) - \frac{1}{\beta}\log Z[\boldsymbol{\eta}] \qquad\qquad f_N = -\frac{1}{\tilde{\beta}N}\log \mathrm{Tr}_{\boldsymbol{\eta}}\; e^{-\tilde{\beta}\ H(\boldsymbol{\eta})} \tag{4}$$

We now make a specific choice for the set C of allowed single-monomer angles, viz. $C = \{(2k{+}1)\pi/q,\ k = 0,\ldots,q{-}1\}$ (q possible orientations per monomer), and we choose the simplest non-trivial type of functionality potential $V(\boldsymbol{\eta}) = \sum_{i=1}^{N}\mu_i\eta_i$. In addition we switch to the Ising version of (4), i.e. $\eta_i \in \{-1,1\}$ for all i (as opposed to $\eta_i \in \mathrm{I\!R}$)§. In combination, studying the stationary state at the largest (genetic) time-scale has now been reduced to the calculation of the free energy per monomer f_N:

$$f_N = -\frac{1}{\tilde{\beta}N}\log\sum_{\boldsymbol{\eta}}\left[\sum_{\boldsymbol{\phi}} e^{-\beta H(\boldsymbol{\phi},\boldsymbol{\eta})}\right]^{\frac{\tilde{\beta}}{\beta}} e^{-\tilde{\beta}\boldsymbol{\mu}\cdot\boldsymbol{\eta}} \qquad\qquad H(\boldsymbol{\phi},\boldsymbol{\eta}) = -\frac{J}{N}\sum_{ij}\eta_i\eta_j\ \delta_{\phi_i,\phi_j} \tag{5}$$

with $\boldsymbol{\eta} \in \{-1,1\}^N$ and with $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$.

The energetic forces in our model favor species-based monomer separation (due to the fast process) and species selection controlled by the disorder variables $\{\mu_i\}$ (due to the slow process). The ease with which these two aims can be achieved simultaneously depends on the statistics of the single-site genetic forces $\{\mu_i\}$; the relative importance of the two objectives is controlled by the ratio $\tilde{\beta}/\beta$ of associated noise levels. Entropic forces act in the opposite direction, favoring randomly distributed monomer orientations and species. The equilibrium state of the system must therefore be characterized by disorder-averaged order parameters which measure the overall joint distribution of orientation- and species variables along the chain.

2.2. Solution via the finite-n replica method

Expression (5), which is of the typical form found for systems with disparate time-scales, can be evaluated using the replica method. The temperature ratio $n = \tilde{\beta}/\beta$ is first regarded as integer, which allows us to write $Z^n[\boldsymbol{\eta}]$ as the partition function of an n-fold replicated system, followed by analytic continuation to non-integer n later (if needed). In the special case $n \to 0$ the polarity numbers $\{\eta_i\}$ reduce to quenched disorder variables, and we return to a simplified version of the model in [17]. For $n = 1$ one recovers the annealed case (although the time-scales remain completely disparate). The limit $n \to \infty$ corresponds to dominant coupling of the two processes: here the polarity dynamics is fully deterministic. When referring to temperature in the remainder of this paper we will mean T, with $\tilde{T}$ linked to T via $\tilde{T} = T/n$.

Insertion into (5) of $\sum_{\phi'}\delta_{\phi_i{}^{\alpha},\phi'} = 1$ followed by exponent linearization via Gaussian integrals gives

$$\begin{aligned} f_N =& -\frac{1}{\tilde{\beta}N}\log\sum_{\boldsymbol{\eta}}\sum_{\boldsymbol{\phi}^1\ldots\boldsymbol{\phi}^n} e^{\frac{\beta J}{N}\sum_{\phi'}\sum_{\alpha=1}^{n}\left[\sum_i \eta_i\delta_{\phi',\phi_i^\alpha}\right]^2 - \tilde{\beta}\sum_i \mu_i\eta_i} \\ =& -\frac{1}{\tilde{\beta}N}\log\int[\prod_{\phi\alpha} dz_{\phi^\alpha}]e^{-\frac{N}{4\beta J}\sum_{\phi\alpha}(z_\phi^\alpha)^2}\sum_{\boldsymbol{\phi}^1\ldots\boldsymbol{\phi}^n}\sum_{\boldsymbol{\eta}} e^{\sum_i \eta_i\left[\sum_\alpha z^\alpha_{\phi_i^\alpha} - \tilde{\beta}\mu_i\right]} \end{aligned}$$

We carry out the summations over the polarity variables (first) and over the replicated angles, and take the thermodynamic limit. This leads us to the following result for the asymptotic free energy per monomer $f = \lim_{N\to\infty} f_N$, which is evaluated by steepest descent:

$$\begin{aligned} \tilde{\beta}f \ =& \ -\lim_{N\to\infty}\frac{1}{N}\log\int[\prod_{\phi\alpha} dz_\phi^\alpha]e^{-\frac{N}{4\beta J}\sum_{\phi\alpha}(z_\phi^\alpha)^2 + \sum_i \log\sum_{\boldsymbol{\phi}} 2\cosh\left[\sum_\alpha z^\alpha_{\phi_\alpha} - \tilde{\beta}\mu_i\right]} \\ =& \ \mathrm{extr}_{\{z_\phi^\alpha\}}\left\{\frac{1}{4\beta J}\sum_{\phi\alpha}(z_\phi^\alpha)^2 - \langle\log\left[e^{-\tilde{\beta}\mu}\prod_\alpha(\sum_\phi e^{z_\phi^\alpha}) + e^{\tilde{\beta}\mu}\prod_\alpha(\sum_\phi e^{-z_\phi^\alpha})\right]\rangle_\mu\right\} \end{aligned} \tag{6}$$

§ Note that this implies that the genetic dynamics (3) will be replaced by either a Glauber-type dynamics, or by a 'soft-spin' Langevin equation (involving the familiar parametrized double-well potential for each η_i) from which the binary version of the model can be obtained as a limit.

in which $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_n)\in C^n$ and $\langle g(\mu)\rangle_\mu=\int\! d\mu\; P(\mu)g(\mu)$, with $P(\mu)=\lim_{N\to\infty}N^{-1}\sum_i\delta[\mu-\mu_i]$. The location $\{Z_\phi^\alpha\}$ of the extremum in (6) is to be calculated by solving the following non-linear saddle-point equations:

$$Z_\psi^\gamma=2\beta J\left\langle\frac{e^{Z_\psi^\gamma-\beta n\mu}\prod_{\alpha\neq\gamma}(\sum_\phi e^{Z_\phi^\alpha})-e^{\beta n\mu-Z_\psi^\gamma}\prod_{\alpha\neq\gamma}(\sum_\phi e^{-Z_\phi^\alpha})}{e^{-\beta n\mu}\prod_\alpha(\sum_\phi e^{Z_\phi^\alpha})+e^{\beta n\mu}\prod_\alpha(\sum_\phi e^{-Z_\phi^\alpha})}\right\rangle_\mu \tag{7}$$

The physical meaning of the order parameters $\{Z_\phi^\alpha\}$ can be determined by adding generating terms to the Hamiltonian (1), which measure the overall polarity at given positions ϕ:

$$H(\boldsymbol{\phi},\boldsymbol{\eta})\to H(\boldsymbol{\phi},\boldsymbol{\eta})+\sum_\phi\chi_\phi L_\phi(\boldsymbol{\phi},\boldsymbol{\eta})\qquad\qquad L_\phi(\boldsymbol{\phi},\boldsymbol{\eta})=\frac{2}{N}\sum_{i=1}^N\eta_i\delta_{\phi_i,\phi} \tag{8}$$

Upon working out the general identity $\lim_{\chi_\phi\to0}\partial f/\partial\chi_\phi=L_\phi=\overline{\langle L_\phi(\boldsymbol{\phi},\boldsymbol{\eta})\rangle}$, with $\langle f(\boldsymbol{\eta},\boldsymbol{\phi})\rangle$ denoting conformational equilibrium averages for fixed $\{\eta_i\}$ and $\overline{f(\boldsymbol{\eta})}$ denoting polarity equilibrium averages, it follows that

$$\frac{1}{n}\sum_\alpha Z_\phi^\alpha=\beta JL_\phi\qquad\qquad L_\phi=\lim_{N\to\infty}\overline{\langle L_\phi(\boldsymbol{\phi},\boldsymbol{\eta})\rangle} \tag{9}$$

Thus L_ϕ is proportional to the disorder-averaged equilibrium expectation value of the average polarity of those monomers which are oriented to angle ϕ. For future use we will also define the overall average equilibrium polarity p:

$$p=\lim_{N\to\infty}\frac{1}{N}\sum_i\overline{\eta_i}=\frac{1}{2}\sum_\phi L_\phi \tag{10}$$

2.3. The replica symmetric solution

In Appendix A we show that the replica-symmetric (RS) solution of our saddle-point equations (7), where $Z_\phi^\alpha=\beta JL_\phi$ for all α, is locally stable against replica-symmetry breaking fluctuations. For such RS solutions one finds the (6) and (7) reducing to, respectively:

$$f_{\rm RS}=\min_{\{\ell_\phi\}}\left\{\frac{J}{4}\sum_\phi\ell_\phi^2-\frac{1}{\beta n}\langle\log\left[(\sum_\phi e^{\beta[J\ell_\phi-\mu]})^n+(\sum_\phi e^{\beta[\mu-J\ell_\phi]})^n\right]\rangle_\mu\right\} \tag{11}$$

$$L_\psi=2\left\langle\frac{e^{\beta[JL_\psi-\mu]}(\sum_\phi e^{\beta[JL_\phi-\mu]})^{n-1}-e^{\beta[\mu-JL_\psi]}(\sum_\phi e^{\beta[\mu-JL_\phi]})^{n-1}}{(\sum_\phi e^{\beta[JL_\phi-\mu]})^n+(\sum_\phi e^{\beta[\mu-JL_\phi]})^n}\right\rangle_\mu \tag{12}$$

Summation over ϕ in (12) gives an alternative expression for the RS average polarity p (10):

$$p=\left\langle\frac{(\sum_\phi e^{\beta[JL_\phi-\mu]})^n-(\sum_\phi e^{\beta[\mu-JL_\phi]})^n}{(\sum_\phi e^{\beta[JL_\phi-\mu]})^n+(\sum_\phi e^{\beta[\mu-JL_\phi]})^n}\right\rangle_\mu \tag{13}$$

Equation (13) allows us to write (12) in the form

$$L_\psi=(1+p)\frac{e^{\beta JL_\psi}}{\sum_\phi e^{\beta JL_\phi}}-(1-p)\frac{e^{-\beta JL_\psi}}{\sum_\phi e^{-\beta JL_\phi}} \tag{14}$$

This latter expression (14) was also found in [17]; however, here the average polarity p is an order parameter, to be solved simultaneously with the $\{L_\phi\}$ from (13), whereas in [17] it was a fixed control parameter.

Since we have shown replica symmetry to be locally stable, for any choice of model parameters, we will henceforth restrict ourselves to RS saddle-points only (there is no evidence for discontinuous RSB transitions).

3. General analytical results

3.1. The high-temperature state

Let us first identify the high-temperature state. Expansion of the free energy (11) for fixed $n > 0$ gives $f_{\rm RS} = -\frac{1}{2}\beta n\langle\mu^2\rangle_\mu - \frac{1}{\beta}\log q - \frac{1}{\beta n}\log 2 + \frac{1}{2}J\ {\rm extr}_{\{\ell_\phi\}}\Phi[\{\ell_\phi\}] + \mathcal{O}(\beta^2)$, with (using $\sum_\phi 1 = q$):

$$\Phi[\{\ell_\phi\}] = \frac{1}{2}\sum_\phi \ell_\phi^2 + \frac{2\beta n}{q}\langle\mu\rangle_\mu \sum_\phi \ell_\phi - \beta J\left[\frac{1}{q}\sum_\phi \ell_\phi^2 + (n-1)(\frac{1}{q}\sum_\phi \ell_\phi)^2\right]$$

It follows that the saddle-point of f is of the form

$$L_\phi = -\frac{2\beta n}{q}\langle\mu\rangle_\mu + \mathcal{O}(\beta^2) \qquad (\beta\to 0)$$

For sufficiently high temperatures the order parameters L_ϕ are thus independent of ϕ. This was to be expected in view of the invariance of (11) under arbitrary permutations of the available monomer orientations $\phi \in C$.

Insertion of the symmetric ansatz $L_\phi = L\ \forall\phi$ (which is a saddle-point of $f_{\rm RS}$ at any temperature) directly into (11,12) gives

$$f_{\rm RS}^{\rm sym} = {\rm min}_\ell \left\{\frac{Jq}{4}\ell^2 - \frac{1}{\beta n}\langle\log\left[e^{\beta n[J\ell-\mu]} + e^{\beta n[\mu - J\ell]}\right]\rangle_\mu\right\} - \frac{1}{\beta}\log q \tag{15}$$

$$L = \frac{2}{q}\langle\tanh[\beta n(JL-\mu)]\rangle_\mu \tag{16}$$

In terms of the average equilibrium polarity (10), which here reduces to $p = qL/2$, equation (16) can be written alternatively as $p = \langle\tanh[\beta n(2Jp/q-\mu)]\rangle_\mu$. The instance where this RS high-temperature state becomes locally unstable can be inferred from the results of Appendix A. In particular (A.7) gives the condition for a zero eigenvalue of the RS Hessian as

$$\det\left[nq(\hat{\boldsymbol{K}} - \hat{\boldsymbol{L}}) + q(\hat{\boldsymbol{L}} - \hat{\boldsymbol{M}}) + \frac{q}{2\beta J}\mathbb{1}\right] = 0$$

Working out the various matrices for the symmetric state $L_\phi = L$, with (16), gives

$$\hat{K}_{\phi\psi} = q^{-2}\langle\tanh^2[n\beta(JL-\mu)]\rangle_\mu \qquad \hat{L}_{\phi\psi} = q^{-2} \qquad \hat{M}_{\phi\psi} = q^{-1}\delta_{\phi\psi}$$

In the symmetric state the Hessian has two distinct eigenvalues, one relating to changes in the amplitude of the symmetric state, and one relating to changes orthogonal to the symmetric state, with associated stability conditions:

locally stable in non–symmetric directions : $qT/2J > 1$
locally stable in symmetric direction : $qT/2J > n\langle 1-\tanh^2[n\beta(JL-\mu)]\rangle_\mu$

For $n \leq 1$ it immediately follows that, as we lower the temperature, the first stability condition is always violated before the second can be. Violation of the second condition implies destabilization in favour of an alternative symmetric solution (i.e. the creation of an alternative solution of equation (16)). Thus symmetric states $L_\phi = L\ \forall\phi$ defined by (15) become locally unstable against symmetry-breaking fluctuations at

$$T_c = 2J/q \tag{17}$$

Unless preceded by a first order transition (which will also turn out to be possible in certain parameter regimes), this describes a second-order phase transition from a non-separated state into one where the hydrophobic and hydrophilic monomers start to prefer distinct orientations. Within the context of our model, the former is to be regarded as a swollen state (S) for the polymer, and the second as a compact state (C). Note that expression (17), which is identical to that found in [17], is independent of n.

3.2. The replica-symmetric ground state for $n > 0$

We now study the limit $\beta \to \infty$ and calculate the ground state of our system, for replica-symmetric solutions and fixed $n > 0$. We define $\ell_+ = \max_\phi \ell_\phi$, $\ell_- = \min_\phi \ell_\phi$ (so $\ell_+ \geq \ell_-$) and the number of locations ϕ for which $\ell_\phi = \ell_\pm$ by $q\pm$, respectively (so $q\pm \geq 1$). The remaining intermediate values for ℓ_ϕ define the set $U = \{\phi | \ \ell_- < \ell_\phi < \ell_+\}$. Similarly we define $L_\pm$ as the values of $\ell_\pm$ at the relevant saddle-point. The ground state is the solution of the following minimization problem:

$$\begin{aligned}
\frac{E_0}{J} &= \min_{\{\ell_\phi\}} \left\{ \frac{1}{4}\sum_\phi \ell_\phi^2 - \lim_{\beta\to\infty} \frac{1}{\beta J n} \langle \ \log \left[(\sum_\phi e^{\beta[J\ell_\phi - \mu]})^n + (\sum_\phi e^{\beta[\mu - J\ell_\phi]})^n \right] \rangle_\mu \right\} \\
&= \min_{\{\ell_\phi\}} \left\{ \frac{1}{4}\sum_\phi \ell_\phi^2 - \frac{1}{J} \langle \max \left[\lim_{\beta\to\infty} \frac{1}{\beta} \log(\sum_\phi e^{\beta[J\ell_\phi - \mu]}), \lim_{\beta\to\infty} \frac{1}{\beta} \log(\sum_\phi e^{\beta[\mu - J\ell_\phi]}) \right] \rangle_\mu \right\} \\
&= \min_{\{\ell_\phi\}} \left\{ \frac{1}{4}\sum_\phi \ell_\phi^2 - \langle \ \max\left[\ell_+ - \mu/J, \ \mu/J - \ell_-\right] \ \rangle_\mu \right\}
\end{aligned}$$

Suppose first that $L_+ = L_-$, which implies $L_\phi = L = 2p/q$ for all ϕ (i.e. a fully symmetric ground state). Now

$$E_0/J = \min_{\{p\}} \left\{ p^2/q - \langle \ |2p/q - \frac{\mu}{J}| \ \rangle_\mu \right\}$$

Next we consider the case $L_+ > L_-$. Minimization with respect to those ℓ_ϕ for which $\phi \in U$ for a given realization of $\{q_+, q_-\}$ reveals that $\ell_\phi = 0$ for $\phi \in U$. We can then minimize further with respect to $\{q_\pm\}$ and find $q_+ = q_- = 1$. Given our previous identification $p = \frac{1}{2}\sum_\phi L_\phi$, here reducing to $p = \frac{1}{2}(L_+ + L_-)$, we write $\ell_\pm = p \pm z$, upon which our problem takes the form

$$E_0/J = \min_{\{p,z\}} \left\{ \frac{1}{2}p^2 + \frac{1}{2}z^2 - z - \langle \ |p - \mu/J| \ \rangle_\mu \right\}$$

We conclude that the minimum corresponds to $x = 1$. Both candidate ground state solutions can thus be expressed in terms of the monotonically increasing function

$$\Lambda(z) = \min_{\{y\}} \left\{ \frac{1}{2} z y^2 - \langle |y - \mu/J| \rangle_\mu \right\}$$

as follows

$$\begin{array}{lll}
\text{symmetric}: & \boldsymbol{L} = L(1,1,\ldots,1,1), & E_0/J = \Lambda(q/2) \\
\text{non-symmetric}: & \boldsymbol{L} = (p-1,0,\ldots,0,p+1), & E_0/J = \Lambda(1) - 1/2
\end{array}$$

Since $q \geq 2$ we may conclude that the ground state is the non-symmetric solution, so

$$E_0/J = \min_{\{p\}} \epsilon(p) \qquad \epsilon(p) = \frac{1}{2}p^2 - \frac{1}{2} - \langle \ |p - \mu/J| \ \rangle_\mu \tag{18}$$

Note that $\epsilon(p)$ is symmetric as soon as $P(\mu)$ is symmetric. The ground state is one where at most two sites $\phi_\pm$ are occupied by monomers, and there are at most two nonzero values $L_\pm$ for the order parameters L_ϕ.

3.3. The limits $n \to 0$, $n \to 1$ and $n \to \infty$, for arbitrary T

If we take $n \to 0$ in equations (11,12,13) (which implies fully random evolution of the polarity variables η_i) we find, as expected, the $p = 0$ version of the solution for the long-range limit of [17]:

$$\lim_{n\to 0} \left[f_{\rm RS} + \frac{\log 2}{\beta n} \right] = \min_{\{\ell_\phi\}} \left\{ \frac{J}{4}\sum_\phi \ell_\phi^2 - \frac{1}{2\beta}\log(\sum_\phi e^{\beta J \ell_\phi}) - \frac{1}{2\beta}\log(\sum_\phi e^{-\beta J \ell_\phi}) \right\}$$

$$L_\psi = \frac{e^{\beta J L_\psi}}{\sum_\phi e^{\beta J L_\phi}} - \frac{e^{-\beta J L_\psi}}{\sum_\phi e^{-\beta J L_\phi}}, \qquad p = 0$$

The (constant) contribution $-\log 2/\beta n$ to the free energy per monomer represents the entropy of the polarity variables. Note that this solution is independent of the distribution $P(\mu)$ of the forces μ_i, as it should.

Putting $n\to 1$ effectively reduces the polarity variables to annealed ones, in spite of the disparate time-scales in the model. Now we find (11,13,14) taking the form

$$\lim_{n\to 1} f_{\rm RS} = \min_{\{\ell_\phi\}}\left\{\frac{J}{4}\sum_\phi \ell_\phi^2 - \frac{1}{\beta}\langle \log 2\sum_\phi \cosh[\beta(J\ell_\phi-\mu)]\rangle_\mu\right\}$$

$$L_\psi = (1+p)\frac{e^{\beta J L_\psi}}{\sum_\phi e^{\beta J L_\phi}} - (1-p)\frac{e^{-\beta J L_\psi}}{\sum_\phi e^{-\beta J L_\phi}}, \qquad\qquad p = \left\langle \frac{\sum_\phi \sinh[\beta(JL_\phi-\mu)]}{\sum_\phi \cosh[\beta(JL_\phi-\mu)]}\right\rangle_\mu$$

Somewhat unexpectedly, these equations are still non-trivial, and retain a rich bifurcation phenomenology as we will show later for specific choices of the number q of possible monomer orientations.

Now we turn to $n\to\infty$, i.e. to fully deterministic genetic dynamics. Here the free energy per monomer (11) reduces to

$$\begin{aligned}
\lim_{n\to\infty} f_{\rm RS} &= \min_{\{\ell_\phi\}}\left\{\frac{J}{4}\sum_\phi \ell_\phi^2 - \langle \max\left[\frac{1}{\beta}\log(\sum_\phi e^{\beta J\ell_\phi})-\mu, \mu+\frac{1}{\beta}\log(\sum_\phi e^{-\beta J\ell_\phi})\right]\rangle_\mu\right\} \\
&= \min_{\{\ell_\phi\}}\left\{\frac{J}{4}\sum_\phi \ell_\phi^2 - \frac{1}{2\beta}\log(\sum_\phi e^{\beta J\ell_\phi}) - \frac{1}{2\beta}\log(\sum_\phi e^{-\beta J\ell_\phi})\right. \\
&\qquad\qquad \left. -\left\langle \left|\mu - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J\ell_\phi}}{\sum_\phi e^{-\beta J\ell_\phi}}\right]\right|\right\rangle_\mu\right\}
\end{aligned} \qquad (19)$$

Equation (13), similarly, becomes

$$p = -\left\langle \mathrm{sgn}\left[\mu - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta JL_\phi}}{\sum_\phi e^{-\beta JL_\phi}}\right]\right]\right\rangle_\mu \qquad (20)$$

3.4. Phase diagrams for $q=2$ and $q=3$

In view of our earlier results regarding the high-temperature states (where $\boldsymbol{L}=L(1,1,\ldots,1,1)$) and ground states (where $\boldsymbol{L}=(p-1,0,\ldots,0,p+1)$) it is natural to assume that the solution of our model will never exhibit more than three distinct values for the order parameters $\{L_\phi\}$. Hence we may restrict our analysis to $q\in\{2,3\}$ and expect at most quantitative changes to emerge for $q>3$. In the phase diagrams to be given below for $q\in\{2,3\}$ we focus on transitions marking bifurcations of local minima of the free energy surface as minimized in (11), rather than on thermodynamic transitions. Since all energetic and entropic barriers in our system are extensive, on the time-scales of any experiment (whether real or numerical, and especially in view of the extremely slow genetic process) one would never detect thermodynamic transitions. The system will in practice be found in the local free energy minimum in whose domain of attraction the initial configuration happened to lie.

For $q=2$ we have only two order parameters, which we can write as $L_\pm = p\pm Z$. Insertion into (11) reveals that now the free energy minimization decouples conveniently into

$$f_{\rm RS} = \min_{\{Z\}}\left\{\frac{JZz^2}{2} - \frac{1}{\beta}\log 2\cosh[\beta JZ]\right\} + \min_{\{p\}}\left\{\frac{Jp^2}{2} - \frac{1}{\tilde{\beta}}\langle \log 2\cosh[\tilde{\beta}(Jp-\mu)]\rangle_\mu\right\} \qquad (21)$$

giving the following independent order parameter equations:

$$Z = \tanh[\beta JZ] \qquad\qquad p = \langle \tanh[\tilde{\beta}(Jp-\mu)]\rangle_\mu \qquad (22)$$

The separation process of hydrophobic from hydrophilic monomers (i.e. the folding), measured by Z, disentangles from the species evolution, as measured by p; with the two processes each having their own independent transitions. This is a consequence of the fact that for $q=2$ the orientation variables ϕ_i effectively become Ising spins, so that upon putting $\delta_{\phi_i\phi_j} \to \frac{1}{2}[1+\sigma_i\sigma_j]$, the 'folding' Hamiltonian in (5) reduces to that of a Mattis [21] magnet, from which the variables $\{\eta_i\}$ can be gauged away.

The uniform high-temperature state $L_\phi = L\ \forall\phi$ (where $Z=0$) always destabilizes via a second order phase transition at the Curie-Weiss temperature $T_c = J$, independent of the distribution $P(\mu)$. The phase phenomenology embodied in the equation for p, in contrast, is described by an equation for a mean-field ferro-magnet at inverse temperature $\tilde{\beta}$ and with random external fields, distributed (apart from a minus sign) according to $P(\mu)$. It will be therefore be dependent on the choice made for $P(\mu)$, with invariance of the problem under $p \to -p$ for symmetric force distributions $P(\mu)$.

For $q=3$ one has $\phi \in \{-\frac{2}{3}\pi, 0, \frac{2}{3}\pi\}$, and there will be three order parameters $\{L_\phi\}$. One can no longer map the model onto a Mattis-like system, and one no longer benefits from the resulting decoupling of orientation degrees of freedom from polarity degrees of freedom which occurred for $q=2$. According to (17), the fully symmetric (i.e. swollen) state $L_\phi = L\ \forall\phi$ now destabilizes locally at the n-independent temperature $T_c = 2J/3$. In addition we know the ground state of the $q=3$ system, for several choices of the force distribution $P(\mu)$, see e.g. (24,29). In contrast to the previous situation $q=2$, however, there appear to be no significant‖ further analytical simplifications possible, and to obtain results on compact states and phase transitions in the regime of intermediate temperatures we have to resort mainly to a numerical analysis of our fundamental equations (11,12,13,14). Bifurcations of new solutions of the saddle-point equations are marked by the smallest eigenvalue of the relevant Hessian becoming zero, i.e.

$$\left[n(\hat{\boldsymbol{K}}-\hat{\boldsymbol{L}}) + \hat{\boldsymbol{L}} - \hat{\boldsymbol{M}}\right]\boldsymbol{x} = -\frac{1}{2}\beta J\boldsymbol{x} \tag{23}$$

(see Appendix A), with the 3×3 matrices as defined in (A.3,A.4,A.5). Extensive numerical analysis of the phases in the region $T < 2J/3$, where the swollen state is locally unstable, reveals a highly non-trivial phase phenomenology, with many simultaneously locally stable saddle-points. Giving all details and lines in this regime of compact states would be more distracting than informative. In contrast, we will focus mainly on transitions in the region $T > 2J/3$, where the swollen state is locally stable, but where in spite of this one generally finds enhanced first order transitions to compact states whose presence is a direct consequence of our coupled dynamics of slow and fast processes¶. Note that for $q=3$ we give phase diagrams with βJ along the horizontal axis (as opposed to $\tilde{\beta}J = n\beta J$) because, in contrast to $q=2$, one here no longer finds that the non-trivial transitions can be expressed in terms of just two effective control parameters (βnJ and $\beta n\sigma$ or βnJ).

4. Results for δ-distributed genetic forces

In the simplest case $P(\mu) = \delta[\mu - \overline{\mu}]$ (i.e. $\mu_i = \overline{\mu}$ for all i) the functionality potential $V(\boldsymbol{\eta})$ in the genetic dynamics favors a single polarity type. Since the general dynamical behaviour of our system can be regarded as a stochastic version of the combined $T=0$ (i.e. deterministic) folding process and the $n\to\infty$ (i.e. deterministic) genetic dynamics, we first inspect the limits $T\to0$ and $n\to\infty$.

4.1. The deterministic limits: $T\to0$ and $n\to\infty$

We first calculate the ground state (i.e. deterministic evolution of configurational angles, $T=0$). We find the function $\epsilon(p)$ in (18) reducing to

$$\epsilon(p) = \frac{1}{2}p^2 - \frac{1}{2} - |p - \frac{\overline{\mu}}{J}|$$

Upon working out the derivatives of $\epsilon(p)$ in the different regimes (characterized by different values of $\mathrm{sgn}[p-\overline{\mu}/J]$) one arrives at the following result:

$$E_0 = -J - |\overline{\mu}|, \qquad \boldsymbol{L} = 2p(1,0,\ldots,0), \qquad \left\{\begin{array}{ll} \overline{\mu} > 0: & p=-1 \\ \overline{\mu} < 0: & p=1 \end{array}\right. \tag{24}$$

‖ It will be clear that our equations can still be simplified partially upon making specific *ansätze* for the saddle-point as in [17], such as the uniform (i.e. swollen) state $L_\phi = L\ \forall\phi$ (which has already been studied in detail for arbitrary q), or states of the form $\boldsymbol{L} = (L_1, L_2, L_2)$ (where our saddle-point and stability problems become 2-dimensional). Note that our system is invariant under permutations of the three monomer orientations, so that states like $\boldsymbol{L} = (L_1, L_2, L_2)$ and $\boldsymbol{L} = (L_2, L_2, L_1)$ are equivalent.

¶ In the long-range limit of [17] one can also find discontinuous bifurcations for very specific values of the control parameters (marking the creation a locally stable state with relatively high free energy, and hence non-thermodynamic), and very close to the general second order transition at $T_c = 2J/q$; this was missed in [17].

For $\overline{\mu} > 0$ the ground state is one where *all* monomers have become hydrophobic, and are found at exactly the same location relative to the chain. For $\overline{\mu} < 0$ the ground state is one where *all* monomers have become hydrophilic, and are again all oriented at the same location. However, one finds also that for $|\overline{\mu}| < J$ the solutions $p = \pm 1$ are both local minima of $\epsilon(p)$ (although the state $p = \mathrm{sgn}[\overline{\mu}]$ will have an energy higher than the ground state $p = -\,\mathrm{sgn}[\overline{\mu}]$ as long as $\overline{\mu} \neq 0$). This will cause remanence effects in the low temperature dynamics. For $\overline{\mu} = 0$ both single-species states $p = \pm 1$ give equivalent (local and global) minima.

Next we turn to $n \to \infty$, i.e. deterministic genetic dynamics. Equation (13), together with the order-parameter equation (14), give us three candidate solution classes, $p \in \{-1, 0, 1\}$ for $n \to \infty$. Only $p = \pm 1$ will be potentially stable, separated by the unstable fixed-point $p = 0$:

$$p = 1: \qquad L_\psi = \frac{2e^{\beta J L_\psi}}{\sum_\phi e^{\beta J L_\phi}}, \qquad\qquad \overline{\mu} < \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]$$

$$\lim_{n\to\infty} f_{\mathrm{RS}} = \frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{\beta}\log(\sum_\phi e^{\beta J L_\phi})$$

$$p = -1: \qquad L_\psi = \frac{-2e^{-\beta J L_\psi}}{\sum_\phi e^{-\beta J L_\phi}}, \qquad\qquad \overline{\mu} > \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]$$

$$\lim_{n\to\infty} f_{\mathrm{RS}} = \frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{\beta}\log(\sum_\phi e^{-\beta J L_\phi})$$

For the swollen state $L_\phi = 2p/q\ \forall\phi$ this implies the following: for $|\overline{\mu}| > 2J/q$ there is only the solution $p = -\,\mathrm{sgn}[\overline{\mu}]$, for $|\overline{\mu}| < 2J/q$ both solutions $p = \pm 1$ are locally stable (but with the lowest free energy obtained for $p = -\,\mathrm{sgn}[\overline{\mu}]$).

In both cases the system evolves towards a single-species state, with either all hydrophilic monomers (the preferred option when $\overline{\mu} < 0$) or all hydrophobic monomers (the preferred option when $\overline{\mu} > 0$). For small $|\overline{\mu}|$ only the preferred states are locally stable, but for sufficiently large $|\overline{\mu}|$ both single-species states are. This behaviour, which is not desirable from a biological point of view, results from the simple fact that for $P(\mu) = \delta[\mu - \overline{\mu}]$ the single-species state with appropriate polarity sign is energetically favorable to both the folding process and the genetic selection process.

4.2. Phase diagrams for $q = 2$

According to (22), for $P(\mu) = \delta[\mu - \overline{\mu}]$ we find the simple Curie-Weiss equation $p = \tanh[\tilde{\beta}(Jp - \overline{\mu})]$. Upon inverting this to $\tilde{\beta}\overline{\mu} = \tilde{\beta}Jp - \frac{1}{2}\log[(1+p)/(1-p)]$ and calculating the derivative with respect to p, one can work out the bifurcation properties of the solution:

$$\begin{array}{ll} \tilde{\beta}J > 1 \ \text{ and }\ |\overline{\mu}| < \mu_c: & \text{two local minima } p \ \text{ in (21)} \\ \text{elsewhere}: & \text{one local minimum} \end{array}$$

with

$$\tilde{\beta}\mu_c = \sqrt{\tilde{\beta}J}\sqrt{\tilde{\beta}J-1} - \frac{1}{2}\log\left[\frac{\sqrt{\tilde{\beta}J}+\sqrt{\tilde{\beta}J-1}}{\sqrt{\tilde{\beta}J}-\sqrt{\tilde{\beta}J-1}}\right] \tag{25}$$

The transitions at $\mu = \pm\mu_c$ are first-order, except for the common point $(\tilde{\beta}J, \tilde{\beta}\overline{\mu}) = (1, 0)$ where they become second-order. In the region of multiple local minima, for $\overline{\mu} \neq 0$ the global minimum has $\mathrm{sgn}[p] = -\,\mathrm{sgn}[\overline{\mu}]$, whereas for $\overline{\mu} = 0$ the local minima for $\tilde{\beta}J > 1$ are equivalent. We note

$$\tilde{\beta}\mu_c = \frac{1}{2}(\tilde{\beta}J-1)^{3/2} + \mathcal{O}((\tilde{\beta}J-1)^2) \qquad (\tilde{\beta}J \to 1)$$

$$\tilde{\beta}\mu_c = \tilde{\beta}J + \mathcal{O}(\log(\tilde{\beta}J)) \qquad (\tilde{\beta}J \to \infty)$$

This agrees with our results regarding the ground state. We have now determined all phases and transitions: there is one second order transition at $\beta J = 1$ from a swollen state with uniformly distributed monomer orientations to a compact state with separation of polarity types, and two first-order transitions at $\overline{\mu} = \pm\mu_c(\beta, n, J)$ (in the region $\beta n J > 1$) marking the creation of multiple locally stable values for the average polarity. These lines are shown in the $(\tilde{\beta}J, \tilde{\beta}\overline{\mu})$ phase diagram, in figure 1. We denote swollen phases with ℓ

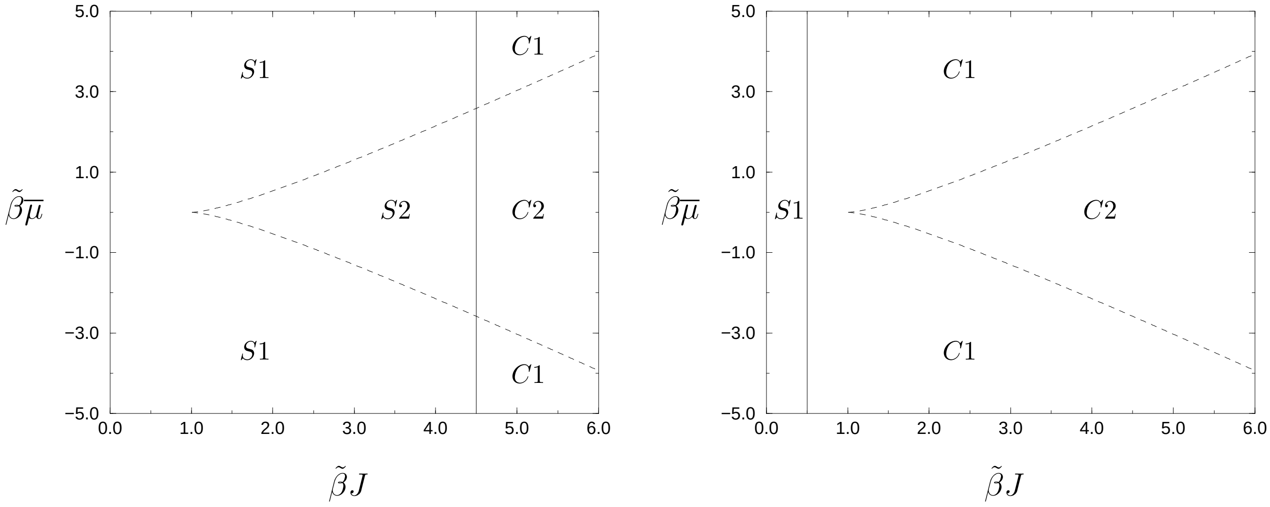

Figure 1. Phase diagrams for $q = 2$ and $P(\mu) = \delta[\mu - \overline{\mu}]$. The four possible phases are $\{S1, S2\}$ (swollen phases with 1 or 2 possible locally stable values of p) and $\{C1, C2\}$ (compact phases with 1 or 2 possible locally stable values of p). The $S1 \to C1$ and $S2 \to C2$ transitions (at $\tilde{\beta}J = n$) are second order (they mark the creation of $Z \neq 0$ solutions of (22)). The $S1 \to S2$ and $C1 \to C2$ transitions are first order (except at $\overline{\mu} = 0$, where the transition is second order). Left diagram: $n = 5$. In this and subsequent phase diagrams all first and second order transitions are drawn as dashed and solid curves, respectively. Right diagram: $n = \frac{1}{2}$. Note that $\tilde{\beta} = \beta n$ and that phase $S2$ exists only for $n > 1$.

possible locally stable values of p as S_ℓ, and compact phases with ℓ possible locally stable values of p as C_ℓ. The number of possible phases depends explicitly on the value of n, since phase $S2$ exists only for $n > 1$.

4.3. Phase diagrams for $q = 3$

Since for both $T \to 0$ and $T \to \infty$ there are only two non-identical values for the order parameters L_ϕ, we expect the solution to be of the form $\boldsymbol{L} = (L_1, L_2, L_2)$ at all temperatures. We thus insert the ansatz $\{L_{\pm 2\pi/3} = \frac{1}{3}(2p + Z), L_0 = \frac{1}{3}(2p - 2Z)\}$ into our saddle-point equations and find

$$p = \frac{e^{-\beta n[\overline{\mu} - \frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ} + e^{-\frac{2}{3}\beta JZ})^n - e^{\beta n[\overline{\mu} - \frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ} + e^{\frac{2}{3}\beta JZ})^n}{e^{-\beta n[\overline{\mu} - \frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ} + e^{-\frac{2}{3}\beta JZ})^n + e^{\beta n[\overline{\mu} - \frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ} + e^{\frac{2}{3}\beta JZ})^n} \tag{26}$$

$$Z = F[Z; p] \qquad\qquad F[Z; p] = \frac{2p[1 - \cosh(\beta JZ)] + 6\sinh(\beta JZ)}{5 + 4\cosh(\beta JZ)} \tag{27}$$

with a corresponding simplification of the bifurcation condition (23). For $Z = 0$ these equations bring us back to the swollen state $L_\phi = \frac{2}{3}p$, with $p = \tanh[\beta n(\frac{2}{3}Jp - \overline{\mu})]$. Comparison with the corresponding equation for $q = 2$ shows that the properties of the swollen state with $q = 3$ can be obtained from those derived for $q = 2$ via the substitution $J \to \frac{2}{3}J$. This gives

$$\begin{array}{lll} \text{swollen state :} & \tilde{\beta}J > \dfrac{3}{2} \;\text{ and }\; |\overline{\mu}| < \mu_c : & \text{two local minima } p \\ & \qquad\text{elsewhere :} & \text{one local minimum} \end{array}$$

with

$$\tilde{\beta}\mu_c = \sqrt{\frac{2}{3}\tilde{\beta}J}\sqrt{\frac{2}{3}\tilde{\beta}J - 1} - \frac{1}{2}\log\left[\frac{\sqrt{\frac{2}{3}\tilde{\beta}J} + \sqrt{\frac{2}{3}\tilde{\beta}J - 1}}{\sqrt{\frac{2}{3}\tilde{\beta}J} - \sqrt{\frac{2}{3}\tilde{\beta}J - 1}}\right] \tag{28}$$

The transitions at $\overline{\mu} = \pm\mu_c$ are first-order, except for the point $(\tilde{\beta}J, \tilde{\beta}\overline{\mu}) = (\frac{3}{2}, 0)$ where they are second-order. Since the swollen state is locally unstable (against collapsed solutions) for $\beta J > \frac{3}{2}$, the transitions at $\overline{\mu} = \pm\mu_c$

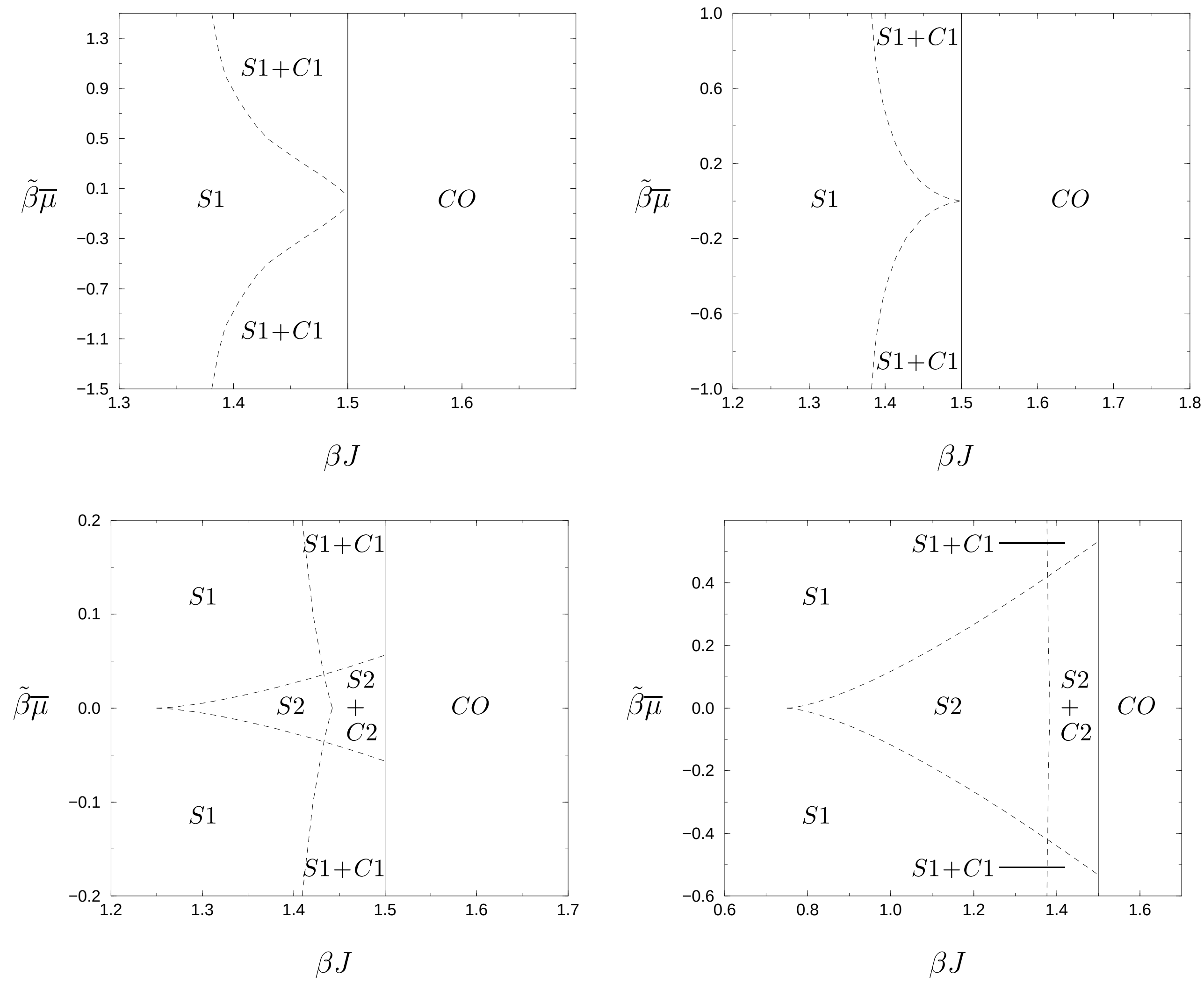

Figure 2. Phase diagrams for $q = 3$ and $P(\mu) = \delta[\mu - \overline{\mu}]$. The possible phases are $\{S\ell\}$ (swollen phases with ℓ possible locally stable values of p), $\{C\ell\}$ (compact phases with ℓ possible locally stable values of p), $\{S\ell{+}C\ell\}$ (phases with 2ℓ possible locally stable states of p, half of which are compact), and CO (compact states only). All transitions are first order, except for the one at $\beta J = \frac{3}{2}$ (which is second order). Upper left diagram: $n = \frac{1}{2}$. Upper right diagram: $n = 1$. Lower left diagram: $n = \frac{6}{5}$. Lower right diagram: $n = 2$.

can be observed only for $n > 1$. It follows from $F[Z;p] = \frac{2}{3}\beta J Z + \mathcal{O}(Z^2)$ that the swollen state, $Z = 0$, indeed destabilizes at $\beta J = \frac{3}{2}$ in favor of a collapsed state of the type above, with $Z \neq 0$.

The main qualitative change observed for $q = 3$ is the emergence of prominent transitions to compact states before the temperature where the swollen state becomes locally unstable. These are in fact also found to be present in the long range limit of [17], where there is no genetic dynamics and where p is a control parameter, but far less prominently and with very small attraction domains of the associated newly created free energy minima. Note that for $n \to 0$ our present model will not exhibit these first order transitions, since this would reduce our system to the $p = 0$ case of [17] (the first order transitions are found only for $p \neq 0$).

5. Results for zero-average binary genetic forces

Our second example is the symmetric binary distribution $P(\mu) = \frac{1}{2}\delta[\mu{+}\sigma] + \frac{1}{2}\delta[\mu{-}\sigma]$, with $\sigma \geq 0$. Here the functionality potential $V(\boldsymbol{\eta})$ in the genetic dynamics discourages the biologically undesirable single-species states with $p = \pm 1$. Again we first inspect the limits $T \to 0$ and $n \to \infty$.

5.1. The deterministic limits: $T \to 0$ and $n \to \infty$

We first calculate the ground state (deterministic evolution of configurational angles: $T = 0$). For the present choice of $P(\mu)$ one finds for the function $\epsilon(p)$ in (18):

$$\epsilon(p) = \frac{1}{2}p^2 - \frac{1}{2} - \frac{1}{2}|p - \frac{\sigma}{J}| - \frac{1}{2}|p + \frac{\sigma}{J}|$$

Upon working out the details in the three regimes $\{p < -\sigma/J,\ |p| < \sigma/J,\ p > \sigma/J\}$ one finds that for $\sigma > 0$ there is always a local minimum of $\epsilon(p)$ at $p = 0$, with $E = -\frac{1}{2}J - \sigma$, and that for $\sigma < J$ there are two additional local minima of $\epsilon(p)$ at $p = \pm 1$ with $E = -J$. The latter will replace $p = 0$ as the global minimum for $\sigma < \frac{1}{2}J$:

$$\begin{array}{llll} \sigma > J/2: & E_0 = -\frac{1}{2}J - \sigma, & p = 0, & \boldsymbol{L} = (-1, 0, \ldots, 0, 1) \\ \sigma < J/2: & E_0 = -J, & p = \pm 1, & \boldsymbol{L} = 2p(1, 0, \ldots, 0) \end{array} \tag{29}$$

For $\sigma > \frac{1}{2}J$ the ground state (29) is one where half of the monomers are hydrophilic and half are hydrophobic (in line with the requirements of the potential $V(\boldsymbol{\eta})$), and where the two species are perfectly separated with all hydrophobic monomers clustered at one site, and all hydrophilic ones clustered at another. The origin of the (undesirable) single-species minima at $p = \pm 1$, for $\sigma < J/2$, can be understood upon realizing that for $\sigma \to 0$ one returns to the δ-distribution studied earlier, with $\overline{\mu} = 0$. For $\sigma = J/2$ both types of states (multiple-species and single-species) give energetically equivalent minima.

For $n \to \infty$ (deterministic genetic dynamics) we get

$$\lim_{n\to\infty} f_{\rm RS} = \min_{\{\ell_\phi\}} \Bigg\{ \frac{J}{4}\sum_\phi \ell_\phi^2 - \frac{1}{2\beta}\log(\sum_\phi e^{\beta J \ell_\phi}) - \frac{1}{2\beta}\log(\sum_\phi e^{-\beta J \ell_\phi}) \\ -\frac{1}{2}\left|\sigma - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J \ell_\phi}}{\sum_\phi e^{-\beta J \ell_\phi}}\right]\right| - \frac{1}{2}\left|\sigma + \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J \ell_\phi}}{\sum_\phi e^{-\beta J \ell_\phi}}\right]\right|\Bigg\} \tag{30}$$

$$p = -\frac{1}{2}\,\text{sgn}\left[\sigma - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]\right] + \frac{1}{2}\,\text{sgn}\left[\sigma + \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]\right] \tag{31}$$

We now have five candidate solutions $p \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$ (due to the symmetry under $\{L_\phi, p\} \to \{-L_\phi, -p\}$ of the minimization problem, the $p = \pm 1$ and $p = \pm\frac{1}{2}$ solutions are mutually equivalent). The potentially stable ones are $p \in \{-1, 0, 1\}$, separated by the unstable fixed-points $p = \pm\frac{1}{2}$:

$$p = 1: \quad L_\psi = \frac{2e^{\beta J L_\psi}}{\sum_\phi e^{\beta J L_\phi}}, \qquad \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right] > \sigma$$

$$\lim_{n\to\infty} f_{\rm RS} = \frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{\beta}\log(\sum_\phi e^{\beta J L_\phi})$$

$$p = 0: \quad L_\psi = \frac{e^{\beta J L_\psi}}{\sum_\phi e^{\beta J L_\phi}} - \frac{e^{-\beta J L_\psi}}{\sum_\phi e^{-\beta J L_\phi}}, \qquad \left|\frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]\right| < \sigma$$

$$\lim_{n\to\infty} f_{\rm RS} = \frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{2\beta}\log(\sum_\phi e^{\beta J L_\phi}) - \frac{1}{2\beta}\log(\sum_\phi e^{-\beta J L_\phi}) - \sigma$$

$$p = -1: \quad L_\psi = \frac{-2e^{-\beta J L_\psi}}{\sum_\phi e^{-\beta J L_\phi}}, \qquad \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right] < -\sigma$$

$$\lim_{n\to\infty} f_{\rm RS} = \frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{\beta}\log(\sum_\phi e^{-\beta J L_\phi})$$

For the swollen state $L_\phi = 2p/q\ \forall\phi$ this implies: for $\sigma > 2J/q$ there is only the solution $p = 0$, for $\sigma < 2J/q$ also the solutions $p = \pm 1$ are locally stable (but with a lower free energy than the $p = 0$ state only when $\sigma < J/q$).

The general picture emerging from the above two limit cases is, as expected, that for sufficiently large σ the single-species states $p = \pm 1$ (which continue to be energetically favorable for the fast process) will indeed be replaced by a biologically more interesting state with equal fractions of hydrophilic and hydrophobic monomers.

5.2. Phase diagrams for $q=2$

Now the order parameter equation becomes

$$p=G(p), \qquad G(p)=\frac{1}{2}\tanh[\tilde{\beta}(Jp{-}\sigma)]+\frac{1}{2}\tanh[\tilde{\beta}(Jp{+}\sigma)] \tag{32}$$

Due to the symmetry $P(\mu)=P(-\mu)$ we always find the state $p=0$ (equal fractions of hydrophobic and hydrophilic monomers). We expand the function $G(p)$ in powers of p,

$$G(p)=\tilde{\beta}J[1-\tanh^2(\tilde{\beta}\sigma)].p-\frac{1}{3}(\tilde{\beta}J)^3\left[1-4\tanh^2(\tilde{\beta}\sigma)+3\tanh^4(\tilde{\beta}\sigma)\right].p^3+\mathcal{O}(p^5)$$

and conclude that the state $p=0$ de-stabilizes when $\tanh^2(\tilde{\beta}\sigma)=1-1/\tilde{\beta}J$, or, equivalently, at

$$\tilde{\beta}\sigma_c=\frac{1}{2}\log\left[\frac{\sqrt{\tilde{\beta}J}+\sqrt{\tilde{\beta}J{-}1}}{\sqrt{\tilde{\beta}J}-\sqrt{\tilde{\beta}J{-}1}}\right] \tag{33}$$

and that this corresponds to a true second-order transition as long as $\tilde{\beta}J<3/2$ (so that $G'''(0)$ is negative). We note that

$$\tilde{\beta}\sigma_c=(\tilde{\beta}J{-}1)^{1/2}+\mathcal{O}(\tilde{\beta}J{-}1) \qquad (\tilde{\beta}J\to 1)$$
$$\tilde{\beta}\sigma_c=\frac{1}{2}\log(\tilde{\beta}J)+\mathcal{O}(1) \qquad (\tilde{\beta}J\to\infty)$$

For $\tilde{\beta}J>3/2$, however, there are first-order transitions away from the state $p=0$ at temperatures higher than the one defined by (33). The locations in the phase diagram of the latter, which start at the triple point $(\tilde{\beta}J,\tilde{\beta}\sigma)=(\frac{3}{2},\text{arctanh}[1/\sqrt{3}])$, are to be solved (numerically) from the coupled transcendental equations $p=G(p)$ and $1=G'(p)$, i.e.

$$p=\frac{1}{2}\tanh[\tilde{\beta}(Jp{-}\sigma)]+\frac{1}{2}\tanh[\tilde{\beta}(Jp{+}\sigma)] \tag{34}$$

$$1=\tilde{\beta}J\left\{1-\frac{1}{2}\tanh^2[\tilde{\beta}(Jp{-}\sigma)]-\frac{1}{2}\tanh^2[\tilde{\beta}(Jp{+}\sigma)]\right\} \tag{35}$$

It is now (in contrast to the case of δ-distributed genetic forces) possible to have *three* locally stable values for the average polarity p (one of which is zero, the other two are different in sign only). This exhausts the phases and transitions: there is one second order transition at $\beta J=1$ from a swollen state with uniformly distributed monomer orientations to a compact state with separation of polarity types, two first-order transition lines (in the region $\beta nJ<\frac{3}{2}$) marking the creation of multiple locally stable values for the average polarity together with de-stabilization of the state $p=0$, and two first order transition lines where two nonzero values of p bifurcate discontinuously (without affecting the stability of $p=0$). These lines are shown in the $(\tilde{\beta}J,\tilde{\beta}\sigma)$ phase diagram, in figure 3. We denote swollen phases with ℓ locally stable values of p as $S\ell$, and compact phases with ℓ locally stable values of p as $C\ell$. The number of possible phases is again found to depend explicitly on n: phase $S2$ exists only for $n>1$, and $S3$ exists only for $n>\frac{3}{2}$.

5.3. Phase diagrams for $q=3$

For the force distribution $P(\mu)=\frac{1}{2}\delta[\mu{-}\sigma]+\frac{1}{2}\delta[\mu{+}\sigma]$ we know the ground state to be $\boldsymbol{L}=(-1,0,1)$ for $\sigma>J/2$ (with $p=0$) and $\boldsymbol{L}=2p(1,0,0)$ for $\sigma<J/2$ (with $p=\pm 1$). We therefore expect the solution to be of the form $\boldsymbol{L}=(L_1,L_2,L_2)$ at all temperatures *at most* when $\sigma<J/2$. Inserting $\{L_{\pm 2\pi/3}=\frac{1}{3}(2p+Z),L_0=\frac{1}{3}(2p-2Z)\}$ into our saddle-point equations now leads to

$$\begin{aligned}p=&\frac{1}{2}\frac{e^{-\beta n[\sigma-\frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ}+e^{-\frac{2}{3}\beta JZ})^n-e^{\beta n[\sigma-\frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ}+e^{\frac{2}{3}\beta JZ})^n}{e^{-\beta n[\sigma-\frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ}+e^{-\frac{2}{3}\beta JZ})^n+e^{\beta n[\sigma-\frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ}+e^{\frac{2}{3}\beta JZ})^n}\\&+\frac{1}{2}\frac{e^{\beta n[\sigma+\frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ}+e^{-\frac{2}{3}\beta JZ})^n-e^{-\beta n[\sigma+\frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ}+e^{\frac{2}{3}\beta JZ})^n}{e^{\beta n[\sigma+\frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ}+e^{-\frac{2}{3}\beta JZ})^n+e^{-\beta n[\sigma+\frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ}+e^{\frac{2}{3}\beta JZ})^n}\end{aligned} \tag{36}$$

$$Z=F[Z;p] \qquad F[Z;p]=\frac{2p[1-\cosh(\beta JZ)]+6\sinh(\beta JZ)}{5+4\cosh(\beta JZ)} \tag{37}$$

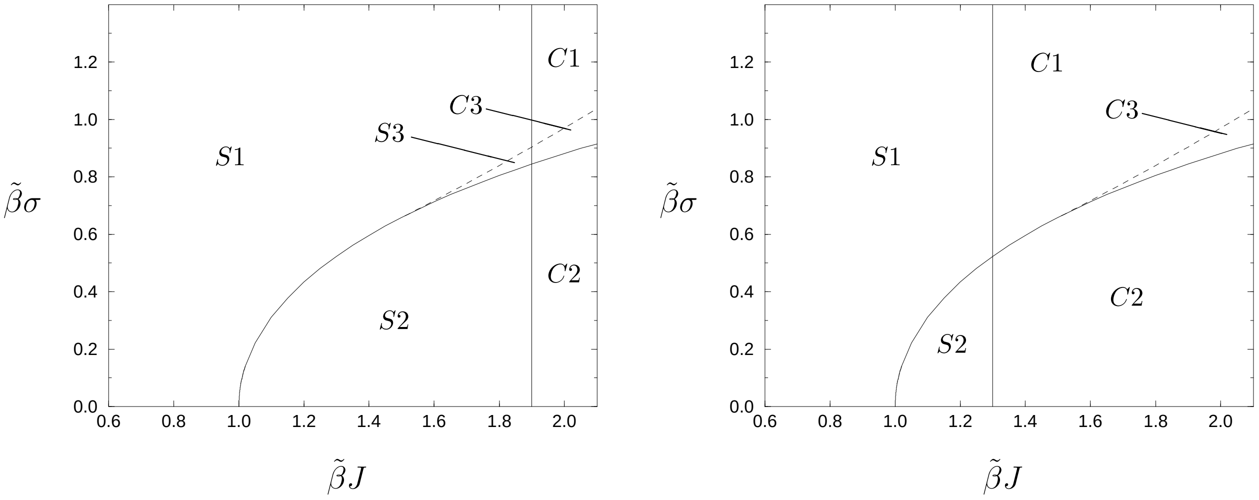

Figure 3. Phase diagrams for $q = 2$ and $P(\mu) = \frac{1}{2}\delta[\mu{-}\sigma] + \frac{1}{2}\delta[\mu{+}\sigma]$. The six possible phases are $\{S1, S2, S3\}$ (swollen phases with 1, 2 or 3 possible locally stable values of p) and $\{C1, C2, C3\}$ (compact phases with 1, 2 or 3 possible locally stable values of p). The $S\ell \to C\ell$ transitions (at $\tilde{\beta}J = n$) as well as the $S1 \to S2$ and $C1 \to C2$ transitions are all second order. The $S1 \to S3$ and $C1 \to C3$ transitions are first order. Left diagram: $n = 5$. Middle diagram: $n = \frac{5}{4}$. Right diagram: $n = \frac{1}{2}$. Note that $\tilde{\beta} = \beta n$, that phase $S2$ exists only for $n > 1$, and that phase $S3$ exists only for $n > \frac{3}{2}$.

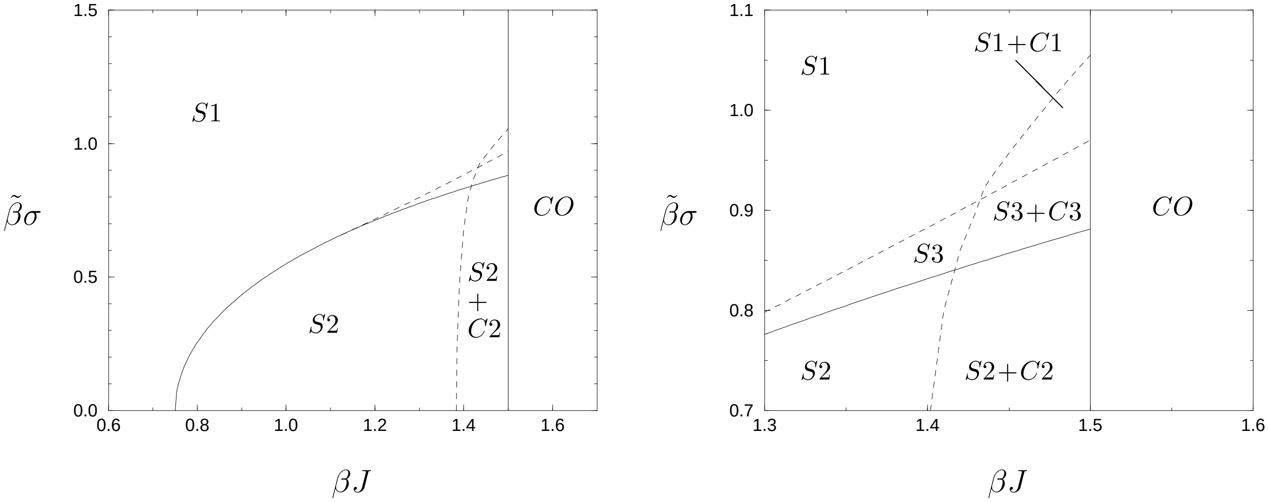

Figure 4. Phase diagrams for $q = 3$ and $P(\mu) = \frac{1}{2}\delta[\mu{-}\sigma] + \frac{1}{2}\delta[\mu{+}\sigma]$, with $n = 2$. The possible phases are $\{S\ell\}$ (swollen phases with ℓ possible locally stable values of p), $\{C\ell\}$ (compact phases with ℓ possible locally stable values of p), $\{S\ell{+}C\ell\}$ (phases with 2ℓ possible locally stable states of p, half of which are compact), and CO (compact states only). First order transitions are indicated by dashed curves, second order ones by solid curves. The right diagram is an enlargement of one of the two areas in the left diagram where all lines nearly meet.

(with a corresponding simplification of the bifurcation condition (23)). For $Z = 0$ these equations bring us back to the swollen state $L_\phi = \frac{2}{3}p$, but now with

$$p = \frac{1}{2}\tanh[\beta n(\frac{2}{3}Jp - \sigma)] + \frac{1}{2}\tanh[\beta n(\frac{2}{3}Jp + \sigma)]$$

Comparison with the corresponding equation for $q = 2$ shows that, for $\sigma < \frac{1}{2}J$, the properties of the swollen state with $q = 3$ can again be obtained from those derived for $q = 2$ via the substitution $J \to \frac{2}{3}J$. This tells us

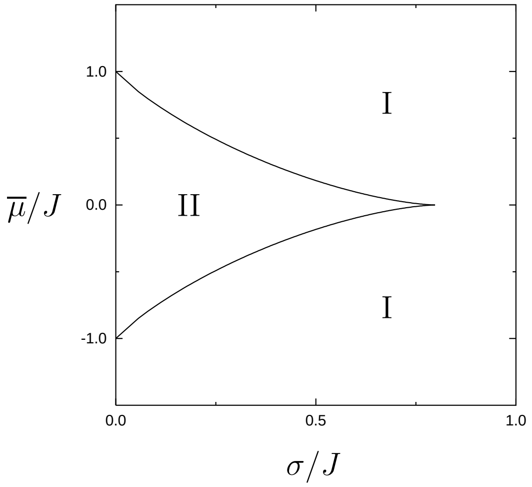

Figure 5. Onset of multiplicity of locally stable states at $T=0$, for Gaussian-distributed forces μ_i (with average $\overline{\mu}$ and variance σ^2). Solid curves: critical lines $\pm\overline{\mu}_c/J$ in parameter space as defined in (42), terminating at $\sigma/J=\sqrt{2/\pi}$. Region I: a single local and global minimum (the ground state), with $\mathrm{sgn}[p]=-\,\mathrm{sgn}[\overline{\mu}]$. Region II: an additional local minimum (with energy higher than the ground state), with $\mathrm{sgn}[p]=\mathrm{sgn}[\overline{\mu}]$.

that the state $p=0$ (which always solves the saddle-point equation) de-stabilizes at

$$\tilde{\beta}\sigma_c=\frac{1}{2}\log\left[\frac{\sqrt{\frac{2}{3}\tilde{\beta}J}+\sqrt{\frac{2}{3}\tilde{\beta}J-1}}{\sqrt{\frac{2}{3}\tilde{\beta}J}-\sqrt{\frac{2}{3}\tilde{\beta}J-1}}\right] \tag{38}$$

and that this corresponds to a true second-order transition as long as $\tilde{\beta}J<9/4$. For $\tilde{\beta}J>9/4$, however, there are first-order transitions away from the state $p=0$ at temperatures higher than the one defined by (38), which start at the triple point $(\tilde{\beta}J,\tilde{\beta}\sigma)=(\frac{9}{4},\mathrm{arctanh}[1/\sqrt{3}])$, to be solved (numerically) from the coupled equations

$$p=\frac{1}{2}\tanh[\tilde{\beta}(\frac{2}{3}Jp-\sigma)]+\frac{1}{2}\tanh[\tilde{\beta}(\frac{2}{3}Jp+\sigma)] \tag{39}$$

$$1=\frac{2}{3}\tilde{\beta}J\left\{1-\frac{1}{2}\tanh^2[\tilde{\beta}(\frac{2}{3}Jp-\sigma)]-\frac{1}{2}\tanh^2[\tilde{\beta}(\frac{2}{3}Jp+\sigma)]\right\} \tag{40}$$

Thus we can have, even for the swollen state, three locally stable values for the average polarity p (one of which is zero, the other two are different in sign only). Since the swollen state is locally unstable (against collapsed solutions) for $\beta J>\frac{3}{2}$, however, the above transitions can be seen only for $n>1$. Since $F[Z;p]=\frac{2}{3}\beta JZ+\mathcal{O}(Z^2)$, the swollen state, $Z=0$ destabilizes at $\beta J=\frac{3}{2}$ in favor of a collapsed state of the type $\boldsymbol{L}=(L_1,L_2,L_2)$.

We show the resulting phase diagram (complemented by numerical analysis of the transitions) in figure 4, for $n=2$. As a result of the symmetry $P(\mu)=P(-\mu)$, here the $S1$ phase always has $p=0$ (in contrast to e.g. the case of δ-distributed forces), whereas $C1$ has $p\neq 0$. The $S2$ and $C2$ phases also involve $p\neq 0$ (with the two possible values different in sign only). The only phase that exists for large values of σ is $S1$. One notes again the crucial dependence on n of the existence of some of the phases (clearly evident in the equations above). It is now clear that upon choosing a sufficiently large value for σ one can eliminate the biologically undesirable single species states at any temperature T.

6. Results for Gaussian-distributed genetic forces

Our third example is the Gaussian distribution $P(\mu)=[2\pi\sigma^2]^{-\frac{1}{2}}e^{-\frac{1}{2}[\mu-\overline{\mu}]^2/\sigma^2}$. Again, provided σ is sufficiently large, the genetic functionality potential discourages the biologically undesirable single-species states. For $\sigma\to\infty$ it follows immediately from (13) that $p=0$.

6.1. *The deterministic limits: $T \to 0$ and $n \to \infty$*

Working out the relevant Gaussian integral in (18) now gives us

$$\epsilon(p) = \frac{1}{2}p^2 - \frac{1}{2} - \frac{\sigma}{J}F\left[\frac{pJ-\overline{\mu}}{\sigma}\right] \qquad\qquad F[x] = x\ \mathrm{Erf}[\frac{x}{\sqrt{2}}] + \sqrt{\frac{2}{\pi}}\ e^{-\frac{1}{2}x^2} \tag{41}$$

It follows that $d\epsilon(p)/dp = p - \mathrm{Erf}[(pJ-\overline{\mu})/\sigma\sqrt{2}]$. Graphical inspection of the functions $\epsilon(p)$ and $d\epsilon(p)/dp$ reveals that for $\overline{\mu} \neq 0$ the global minimum of $\epsilon(p)$ (i.e. the true ground state) is always at a value of p with $\mathrm{sgn}[p] = -\ \mathrm{sgn}[\overline{\mu}]$, but that for sufficiently small σ there will be an additional local minimum (bifurcating discontinuously, together with a corresponding local maximum) with $\mathrm{sgn}[p] = \mathrm{sgn}[\overline{\mu}]$, provided $|\overline{\mu}| < J$. The bifurcation point is calculated by solving simultaneously the equations $d\epsilon(p)/dp = d^2\epsilon(p)/dp^2 = 0$. Working out these equations and eliminating p leads us to the two critical lines $\overline{\mu} = \pm\overline{\mu}_c(\sigma)$, which separate a region with a single local minimum from a region with two local minima:

$$\frac{\overline{\mu}_c}{J} = \mathrm{Erf}\left[\frac{1}{\sqrt{2}}\sqrt{\log\left[\frac{2J^2}{\pi\sigma^2}\right]}\right] - \frac{\sigma}{J}\sqrt{\log\left[\frac{2J^2}{\pi\sigma^2}\right]} \tag{42}$$

The result is shown in figure 5. For $\sigma \to 0$ one finds $\overline{\mu}_c/J \to 1$, and for $\sigma/J \to \sqrt{2/\pi}$ one finds $\overline{\mu}_c \to 0$. Thus for $\sigma \to 0$ we recover the ground state of the example $P(\mu) = \delta[\mu - \overline{\mu}]$, as it should. For $\overline{\mu} = 0$, on the other hand, we have $d\epsilon(p)/dp = p - \mathrm{Erf}[pJ/\sigma\sqrt{2}]$. Now one has only the solution $p = 0$ (equal numbers of hydrophobic and hydrophilic monomers) for $\sigma/J > \sqrt{2/\pi}$, and two equivalent stable solutions $\pm p$ with $\mathrm{sgn}[p] = -\ \mathrm{sgn}[\overline{\mu}]$ for $\sigma/J < \sqrt{2/\pi}$, separated by a second-order transition at $\sigma/J = \sqrt{2/\pi}$.

For $n \to \infty$ (deterministic genetic dynamics one finds, upon doing the required integrals:

$$\lim_{n\to\infty} f_{\mathrm{RS}} = \mathrm{min}_{\{L_\phi\}}\left\{\frac{J}{4}\sum_\phi L_\phi^2 - \frac{1}{2\beta}\log(\sum_\phi e^{\beta J L_\phi}) - \frac{1}{2\beta}\log(\sum_\phi e^{-\beta J L_\phi}) \right.$$
$$\left. -F\left[\frac{1}{\sigma}\left(\overline{\mu} - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]\right)\right]\right\} \tag{43}$$

$$p = -\ \mathrm{Erf}\left[\frac{1}{\sigma\sqrt{2}}\left(\overline{\mu} - \frac{1}{2\beta}\log\left[\frac{\sum_\phi e^{\beta J L_\phi}}{\sum_\phi e^{-\beta J L_\phi}}\right]\right)\right] \tag{44}$$

with the function $F[x]$ defined in (41), and together with our order-parameter equation (14) for the $\{L_\phi\}$. Here one will generally find a continuous dependence of the average polarity p on the system's control parameters. Inspection of the swollen state $L_\phi = 2p/q\ \forall\phi$ here implies studying the saddle-point equation $p = \mathrm{Erf}[(2pJ/q - \overline{\mu})/\sigma\sqrt{2}]$. For $\overline{\mu} = 0$ one finds only the $p = 0$ state for $\sigma/J > 2\sqrt{2}/q\sqrt{\pi}$, which destabilizes at $\sigma/J = 2\sqrt{2}/q\sqrt{\pi}$ in favour of two energetically equivalent $p \neq 0$ solutions (with opposite signs). For $\overline{\mu} \neq 0$ there is always a solution with $\mathrm{sgn}[p] = -\ \mathrm{sgn}[\overline{\mu}]$, which always carries the lowest free energy. A bifurcation analysis of our saddle-point equation for p reveals further that an alternative solution with $\mathrm{sgn}[p] = \mathrm{sgn}[\overline{\mu}]$ is created discontinuously at $\overline{\mu} = \pm\overline{\mu}_c$, where

$$\frac{\overline{\mu}_c}{J} = \frac{2}{q}\ \mathrm{Erf}\left[\sqrt{\log\left(\frac{2\sqrt{2}J}{q\sigma\sqrt{\pi}}\right)}\right] - \frac{\sigma\sqrt{2}}{J}\sqrt{\log\left(\frac{2\sqrt{2}J}{q\sigma\sqrt{\pi}}\right)}$$

(provided $|\overline{\mu}|/J < (2/q)\mathrm{Erf}[\sqrt{2/\pi}]$).

As with binary distributed μ, again we find that for sufficiently large σ the single-species states $p = \pm 1$ are replaced by biologically more interesting states with equal fractions of hydrophilic and hydrophobic monomers.

6.2. *Phase diagrams for $q = 2$*

Working out the $q = 2$ equations (22) for zero-average Gaussian genetic forces (since choosing $\overline{\mu} \neq 0$ always has a deteriorating effect) leads us to the following saddle-point equation for the polarity order parameter p (with the short-hand $Dz = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}z^2}$):

$$p = G(p), \qquad\qquad G(p) = \int Dz\ \tanh[\tilde{\beta}(Jp - \sigma z)] \tag{45}$$

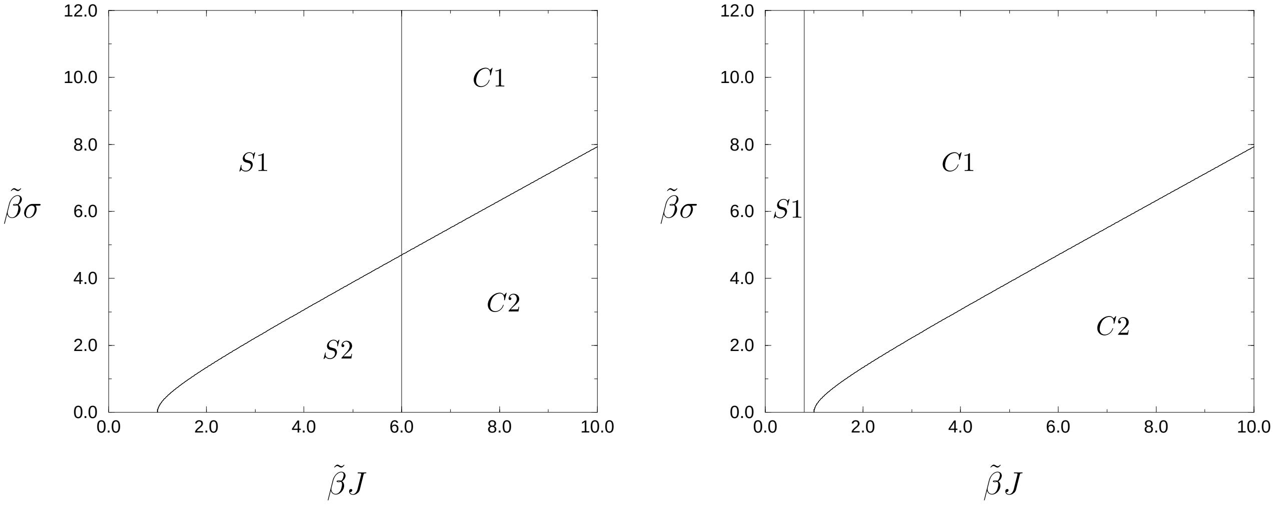

Figure 6. Phase diagrams for $q=2$ and $P(\mu)=[2\pi\sigma^2]^{-1/2}e^{-\frac{1}{2}\mu^2/\sigma^2}$. The four possible phases are $\{S1,S2\}$ (swollen phases with 1 or 2 possible locally stable values of p) and $\{C1,C2\}$ (compact phases with 1 or 2 possible locally stable values of p). Here all transitions are second order. Left diagram: $n=6$. Right diagram: $n=\frac{3}{4}$. Note that $\tilde{\beta}=\beta n$, and that phase $S2$ exists only for $n>1$.

Again $p=0$ is always a solution. We expand $G(p)$ for small p and find

$$G(p)=\tilde{\beta}Jp\int\!Dz\left[1-\tanh^2(\tilde{\beta}\sigma z)\right]$$

$$-\frac{1}{3}(\tilde{\beta}J)^3p^3\int\!Dz\left[1-4\tanh^2(\tilde{\beta}\sigma z)+3\tanh^4(\tilde{\beta}\sigma z)\right]+\mathcal{O}(p^5)$$

and conclude that the state $p=0$ de-stabilizes at

$$1=\tilde{\beta}J\int Dz\left[1-\tanh^2(\tilde{\beta}\sigma z)\right] \tag{46}$$

We note that

$$\tilde{\beta}\sigma_c=(\tilde{\beta}J-1)^{1/2}+\mathcal{O}(\tilde{\beta}J-1)\qquad(\tilde{\beta}J\to 1)$$

$$\tilde{\beta}\sigma_c=\sqrt{\frac{2}{\pi}}\tilde{\beta}J+\mathcal{O}(1)\qquad(\tilde{\beta}J\to\infty)$$

(in line with our earlier results on the ground state). The transition (46) would be preceded by a first order one from the point onwards where $G'''(0)\geq 0$. Along the line (46) this latter condition translates into

$$1\leq\frac{3}{4}\tilde{\beta}J\int\!Dz\left[1-\tanh^4(\tilde{\beta}\sigma z)\right]$$

Numerical analysis reveals that this inequality is never satisfied, and there is no evidence for first order transitions. The final picture for $q=2$ is thus as follows. There are only second order transitions: one occurs at $\beta J=1$ from a state with uniformly distributed monomer orientations (representing a swollen state) to a state with separation of polarity types (representing a compact state), the remaining two occur for $\sigma=\pm\sigma_c$, where σ_c denotes the non-negative solution of (46), and mark the creation of two locally stable non-zero values for the average polarity (which differ in sign only) together with de-stabilization of the state $p=0$. These lines are shown in the $(\tilde{\beta}J,\tilde{\beta}\sigma)$ phase diagram, in figure 6. We again denote swollen phases with ℓ locally stable values of p as $S\ell$, and compact phases with ℓ locally stable values of p as $C\ell$. As in the previous examples, phase $S2$ exists only for $n>1$.

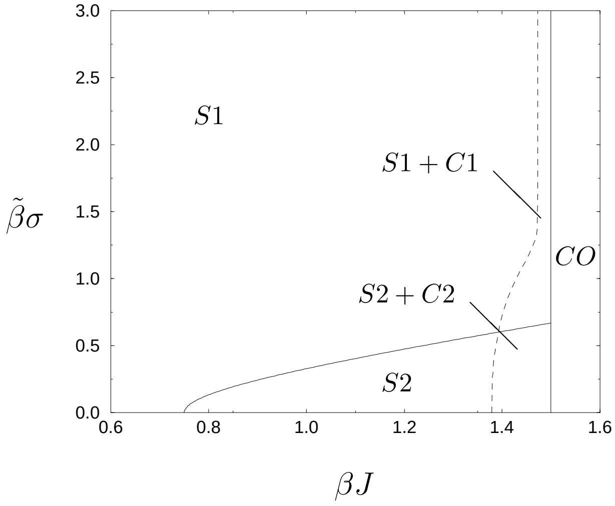

Figure 7. Phase diagram for $q = 3$ and $P(\mu) = [2\pi\sigma^2]^{-\frac{1}{2}} e^{-\frac{1}{2}\mu^2/\sigma^2}$, with $n = 2$. The possible phases are $\{S\ell\}$ (swollen phases with ℓ possible locally stable values of p), $\{C\ell\}$ (compact phases with ℓ possible locally stable values of p), $\{S\ell{+}C\ell\}$ (phases with 2ℓ possible locally stable states of p, half of which are compact), and CO (compact states only). First order transitions are indicated by dashed curves, second order ones by solid curves.

6.3. Phase diagrams for $q = 3$

For the force distribution $P(\mu) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2}\mu^2/\sigma^2}$ we know the ground state to be $\boldsymbol{L} = (-1, 0, 1)$ for $\sigma > J\sqrt{2/\pi}$ (with $p = 0$) and $\boldsymbol{L} = 2p(1, 0, 0)$ for $\sigma < J\sqrt{2/\pi}$ (with $p = \pm 1$). We therefore expect the solution to be of the relatively form $\boldsymbol{L} = (L_1, L_2, L_2)$ at all temperatures *at most* when $\sigma < J/2$. Inserting $\{L_{\pm 2\pi/3} = \frac{1}{3}(2p + Z), L_0 = \frac{1}{3}(2p - 2Z)\}$ into our saddle-point equations now leads to

$$p = \int Du \left[\frac{e^{\beta n[\sigma u + \frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ} + e^{-\frac{2}{3}\beta JZ})^n - e^{-\beta n[\sigma u + \frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ} + e^{\frac{2}{3}\beta JZ})^n}{e^{\beta n[\sigma u + \frac{2}{3}Jp]}(2e^{\frac{1}{3}\beta JZ} + e^{-\frac{2}{3}\beta JZ})^n + e^{-\beta n[\sigma u + \frac{2}{3}Jp]}(2e^{-\frac{1}{3}\beta JZ} + e^{\frac{2}{3}\beta JZ})^n} \right] \tag{47}$$

$$Z = F[Z; p] \qquad\qquad F[Z; p] = \frac{2p[1 - \cosh(\beta JZ)] + 6\sinh(\beta JZ)}{5 + 4\cosh(\beta JZ)} \tag{48}$$

For $Z = 0$ these equations bring us back to the swollen state $L_\phi = \frac{2}{3}p$, but now with

$$p = \int Dz \;\tanh[\beta n(\frac{2}{3}Jp + \sigma z)]$$

Comparison with the corresponding equation for $q = 2$ shows that, for $\sigma < J\sqrt{2/\pi}$, the properties of the swollen state with $q = 3$ can again be obtained from those derived for $q = 2$ via the substitution $J \to \frac{2}{3}J$. This tells us that the state $p = 0$ (which always solves the saddle-point equation) de-stabilizes at

$$1 = \frac{2}{3}\tilde{\beta}J \int Dz \left[1 - \tanh^2(\tilde{\beta}\sigma z)\right] \tag{49}$$

and that this corresponds to a true second-order transition.

We show the resulting phase diagram (complemented by numerical analysis of the transitions) in figure 7, for $n = 2$. As a result of the symmetry $P(\mu) = P(-\mu)$ the $S1$ phase always has $p = 0$ (in contrast to e.g. δ-distributed forces). The $S2$ and $C2$ phases involve $p \neq 0$ (with the two values different in sign only). One notes again the crucial dependence on n of the existence of some of the phases (clearly evident in the equations above). It is clear that upon choosing a sufficiently large value for σ one can eliminate the biologically undesirable single species states at any temperature T.

7. Discussion

Our objective was to define and study a simple solvable model describing the coupled dynamics of (fast) compactification and (slow) genetic monomer sequence selection in hetero-polymers, driven by the competing demands of functionality and reproducibility of the resulting folded structures, as a first step towards understanding the genesis and statistics of natural amino-acid sequences in proteins. Our model is a simple mean-field hetero-polymer whose state is described by two degrees of freedom per monomer i: an angle ϕ_i, giving its orientation relative to the backbone, and a binary variable η_i, giving its polarity (i.e. hydrophobic vs. hydrophilic). The evolution of the orientation variables represents folding; that of the polarity variables (which involves changing monomer species) represents genetic evolution. The latter is assumed to be adiabatically slow compared to the former. There is also explicit (quenched) site disorder in the model, in the form of a simple (site-factorized) random functionality potential for the monomer sequences.

We have solved our model in equilibrium using the finite-n replica formalism, designed to analyze statistical mechanical systems with disparate time-scales, where n denotes the ratio of the noise levels (or temperatures) in the two stochastic processes. Replica symmetry is stable, and we are able to derive closed equations for the system's order parameters, which measure the degree of separation of the monomer species, and can be interpreted in terms of swollen versus compact states. This leads to explicit analytical results for ground states and high-temperature states. Solution of our order-parameter equations at intermediate temperatures requires a combination of analytical and numerical methods. Since the qualitative behaviour of the system can already be captured by restricting ourselves to $q \leq 3$ (q denotes the number of possible orientation angles per monomer), we calculate phase diagrams for $q \in \{2, 3\}$ and for different choices of the disorder in the sequence functionality potential. These exhibit a rich phenomenology of first- and second-order phase transitions, and shed light on the emergence of monomer species statistics (characterized by the fraction of hydrophobic versus hydrophilic monomers) from the underlying genetic dynamics. The properties of the functionality potential are found to have a large impact on the nature of the phase diagram; the distribution of genetic forces needs to be sufficiently broad relative to its average, in order to prevent the system from evolving towards a single-species state (where all monomers have the same polarity). This is perfectly compatible with the efficiency-driven biological need for large sequence diversity (in order to use the available hetero-polymer hardware for as many different biological functions as possible), which again favors the $p = 0$ states.

There is much scope for further work. At a theoretical level one could repeat the above analysis for models in which the folding process also includes short-range forces (e.g. hydrogen bonds and steric forces), as in [17]), which would involve n-replicated transfer matrices, or study the dynamics (at either the fast or the slow time-scale, or both). One could also inspect more realistic sequence functionality potentials $V(\boldsymbol{\eta})$ with a large number of local minima. At the level of increasing biological realism one could introduce more sophisticated degrees of freedom for the monomers, such as three real-valued orientation angles per monomer, or allow for a realistic number of monomer species (as opposed to two) and endow each with additional physical characteristics.

A final criticism to be raised against the present model is that its genetic Hamiltonian favours sequences which allow for efficient shielding of hydrophobic monomers and which meet functionality requirements. Efficient shielding, and hence reliable compactification, is our simplified measure of structure reproducibility. It would be interesting to construct and solve a model in which the genetic Hamiltonian really measures the number of meta-stable states, and rewards sequences which are not only guaranteed to generate a compact conformation (as in the present model) but also a unique one.

References

[1] Research Collaboratory for Structural Bioinformatics (NSF, DoE, NIH) 1987 *Protein Data Bank*, `http://www.pdb.bnl.gov/index.html`

[2] National Biomedical Research Foundation (USA), Munich Information Center for Protein Sequences, Japanese International Protein Sequence Database 1984 *Protein Information Resource*, `http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml`

[3] Brooks C 1998 *Curr. Opinion. Struct. Biol.* **8** 222

[4] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 *Protein Science* **4** 561

[5] Pande V and Roksar D 1999 *Proc. Natl. Acad. Sci. USA* **96** 9062

[6] Wang T, Miller J, Wingreen N S, Tang C and Dill K A 2000 *J. Chem. Phys.* **113** 8329

[7] Garel T, Orland H and Thirumalai D 1996 in *New Developments in Theoretical Studies of Proteins* R Elber Ed. (Singapore: World Scientific)

[8] Garel T, Orland H and Pitard E 1998 in *Spin Glasses and Random Fields* A P Young Ed. (Singapore: World Scientific)

[9] Coolen A C C, Penney R W and Sherrington D 1993 *Phys. Rev. B* **48** 16116

[10] Penney R W, Coolen A C C and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3681

[11] Dotsenko V, Franz S and Mézard M 1994 *J. Phys. A: Math. Gen.* **27** 2351

[12] Penney R W and Sherrington D 1994 *J. Phys. A: Math. Gen.* **27** 4027

[13] Caticha N 1994 *J. Phys. A: Math. Gen.* **27** 5501

[14] Jongen G, Anemüller J, Bollé D, Coolen A C C and Pérez-Vicente C J 2000 *J. Phys. A: Math. Gen.* **34** 3957

[15] Van Mourik J and Coolen A C C 2001 *J. Phys. A: Math. Gen.* **34** L111

[16] Uezu T and Coolen A C C 2002 *J. Phys. A: Math. Gen.* **35** 2761

[17] Skantzos N S, van Mourik J and Coolen A C C 2000 *J. Phys. A: Math. Gen.* **34** 4437

[18] Garel T and Orland H 1988 *Europhys. Lett.* **6** 597

[19] Shakhnovich E I and Gutin A M 1989 *J. Phys. A: Math. Gen.* **22** 1647

[20] Stafos C D, Gutin A M and Shakhnovich E I 1993 *Phys. Rev. E* **48** 465

[21] Mattis D C 1976 *Phys. Lett.* **56A** 421

Appendix A. Stability analysis

General properties

Here we analyse the local stability properties of the saddle-points of the free energy surface per monomer $f[\{z\}]$, which is extremized in (6), by studying the eigenvalue problem of the Hessian $D_{\alpha\phi,\gamma\psi} = \partial^2 f[\{z\}]/\partial z^\alpha_\phi \partial z^\gamma_\psi$. In replica-symmetric saddle-points, where $z^\alpha_\phi = z_\phi = \beta J L_\phi$ for all α, the Hessian takes the following form

$$D_{\alpha\phi,\gamma\psi} = \frac{1}{\tilde{\beta}}\left\{\frac{1}{2\beta J}\delta_{\alpha\gamma}\delta_{\phi\psi} + \tilde{D}_{\alpha\phi,\gamma\psi}\right\} \tag{A.1}$$

$$\tilde{D}_{\alpha\phi,\gamma\psi} = \hat{K}_{\phi\psi} - \hat{L}_{\phi\psi} + \delta_{\alpha\gamma}[\hat{L}_{\phi\psi} - \hat{M}_{\phi\psi}] \tag{A.2}$$

where

$$\hat{K}_{\phi\psi} = \langle [w_+ v^+_\psi - w_- v^-_\psi][w_+ v^+_\phi - w_- v^-_\phi]\rangle_\mu \tag{A.3}$$

$$\hat{L}_{\phi\psi} = \langle w_+ v^+_\psi v^+_\phi + w_- v^-_\psi v^-_\phi\rangle_\mu \tag{A.4}$$

$$\hat{M}_{\phi\psi} = \delta_{\phi\psi}\langle w_+ v^+_\psi + w_- v^-_\psi\rangle_\mu \tag{A.5}$$

and with the short-hands

$$w_\pm = \frac{e^{\mp\tilde{\beta}\mu}(\sum_\phi e^{\pm\beta J L_\phi})^n}{e^{-\tilde{\beta}\mu}(\sum_\phi e^{\beta J L_\phi})^n + e^{\tilde{\beta}\mu}(\sum_\phi e^{-\beta J L_\phi})^n} \qquad v^\pm_\phi = \frac{e^{\pm\beta J L_\phi}}{\sum_{\phi'} e^{\pm\beta J L_{\phi'}}}$$

The three matrices $\{\hat{K}, \hat{L}, \hat{M}\}$ are all positive definite and symmetric, and the non-negative quantities $w_\pm$ and $v^\pm_\phi$ obey the normalization relations $w_+ + w_- = 1$ and $\sum_\phi v^\pm_\phi = 1$. For high temperatures (where we know the system is ergodic) we just find

$$\beta^2 D_{\alpha\phi,\gamma\psi} = \frac{1}{n}\left\{\frac{1}{2J}\delta_{\alpha\gamma}\delta_{\phi\psi} + \mathcal{O}(\beta)\right\} \qquad (\beta \to 0)$$

(i.e. D is positive definite for $\beta \to 0$). Hence the condition for a local de-stabilization of an RS saddle-point $\{L_\phi\}$ is given in terms of the smallest eigenvalue of $\tilde{D}$ as

$$\sum_{\gamma\psi} \tilde{D}_{\alpha\phi,\gamma\psi} x^\gamma_\psi = \lambda x^\alpha_\phi, \qquad \lambda_{\min} = -1/2\beta J \tag{A.6}$$

We now set out to determine the eigenvectors and eigenvalues of (A.2).

RS & RSB eigenvectors, local stability of the RS saddle-point

Replica-symmetric eigenvectors, describing fluctuations within the RS subspace, are of the form $x_\phi^\alpha = x_\phi$ for all α. Insertion into the eigenvalue problem of (A.2) gives

$$\Big[n(\hat{\boldsymbol{K}}-\hat{\boldsymbol{L}})+\hat{\boldsymbol{L}}-\hat{\boldsymbol{M}}\Big]\,\boldsymbol{x}=\lambda_{\rm RS}\;\boldsymbol{x} \tag{A.7}$$

with $\boldsymbol{x}=\{x_\phi\}$. Since the matrices (A.3,A.4,A.5) are $q\times q$ ones, we know that the RS eigenspace is q-dimensional. The matrix $\tilde{D}$ is self-adjoint, so we know that the RSB eigenvectors, which we write as y_ϕ^α, must be orthogonal to the above q-dimensional RS eigenspace; they must obey the q conditions $\sum_\alpha y_\phi^\alpha=0$ for all ϕ. Insertion into the eigenvalue problem of (A.2) now gives

$$\sum_\psi[\hat{L}_{\phi\psi}-\hat{M}_{\phi\psi}]y_\psi^\alpha=\lambda_{\rm RSB}\;y_\phi^\alpha \tag{A.8}$$

Together with the othogonality conditions this leaves an $(n-1)q$ dimensional RSB eigenspace, as it should. We conclude that, in view of (A.6), that second-order transitions of the RS and RSB type occur when, respectively

$$\text{RS transitions}: \qquad \min_{\boldsymbol{x}\in{\rm I\!R}^q,\;\neq\boldsymbol{0}}\left[\frac{\boldsymbol{x}\cdot[n(\hat{\boldsymbol{K}}\!-\!\hat{\boldsymbol{L}})+\hat{\boldsymbol{L}}-\hat{\boldsymbol{M}}]\boldsymbol{x}}{\boldsymbol{x}^2}\right]=-\frac{1}{2\beta J} \tag{A.9}$$

$$\text{RSB transitions}: \qquad \min_{\boldsymbol{x}\in{\rm I\!R}^q,\;\neq\boldsymbol{0}}\left[\frac{\boldsymbol{x}\cdot[\hat{\boldsymbol{L}}-\hat{\boldsymbol{M}}]\boldsymbol{x}}{\boldsymbol{x}^2}\right]=-\frac{1}{2\beta J} \tag{A.10}$$

Below we show that the matrix $\hat{K}-\hat{L}$ is negative definite. As a consequence we can be certain that a destabilization of an RS saddle-point will always occur within the RS eigenspace, since

$$\min_{\boldsymbol{x}\in{\rm I\!R}^q,\;\neq\boldsymbol{0}}\left[\frac{\boldsymbol{x}\cdot[n(\hat{\boldsymbol{K}}\!-\!\hat{\boldsymbol{L}})+\hat{\boldsymbol{L}}-\hat{\boldsymbol{M}}]\boldsymbol{x}}{\boldsymbol{x}^2}\right]<\min_{\boldsymbol{x}\in{\rm I\!R}^q,\;\neq\boldsymbol{0}}\left[\frac{\boldsymbol{x}\cdot[\hat{\boldsymbol{L}}-\hat{\boldsymbol{M}}]\boldsymbol{x}}{\boldsymbol{x}^2}\right]$$

(only non-physical RS saddle-points can be locally unstable against RSB fluctuations). Hence replica-symmetry will always be locally stable. What remains is to confirm that $\hat{K}-\hat{L}$ is indeed negative definite. To do so we first define $\langle x\rangle_\pm=\sum_\phi v_\phi^\pm x_\phi$. This allows us to write

$$\begin{aligned}\sum_{\phi\phi'}x_\phi[\hat{K}_{\phi\phi'}-\hat{L}_{\phi\phi'}]x_{\phi'} &= \Big\langle\;[w_+\langle x\rangle_+-w_-\langle x\rangle_-]^2-w_+\langle x\rangle_+^2-w_-\langle x\rangle_-^2\;\Big\rangle_\mu\\ &\leq \Big\langle\;[w_+|\langle x\rangle_+|+w_-|\langle x\rangle_-|]^2-w_+|\langle x\rangle_+|^2-w_-|\langle x\rangle_-|^2\;\Big\rangle_\mu\\ &\leq 0\end{aligned}$$

This completes our proof.